(1)

AN ASSESSMENT OF THE QUALITY OF SAMPLING PROCEDURES REPORTED IN CLINICAL NURSING RESEARCH: A PILOT STUDY

by

Nancy O'Pry Gentry

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/CI/NR-88- 153	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN ASSESSMENT OF QUALITY OF SAMPLING PROCEDURES REPORTED IN CLINICAL NURSING RESEARCH: A PILOT STUDY		5. TYPE OF REPORT & PERIOD COVERED MS THESIS 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) NANCY O'PRY GENTRY		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS AFIT STUDENT AT: UNIVERSITY OF NORTH CAROLINA - CHAPEL HILL		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE 1988 13. NUMBER OF PAGES 122
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) AFIT/NR Wright-Patterson AFB OH 45433-6583		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION, DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

DISTRIBUTED UNLIMITED: APPROVED FOR PUBLIC RELEASE

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

SAME AS REPORT

18. SUPPLEMENTARY NOTES

Approved for Public Release: IAW AFR 190-1
LYNN E. WOLAVER Lynn Wolaver 21 July 88
Dean for Research and Professional Development
Air Force Institute of Technology
Wright-Patterson AFB OH 45433-6583

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

ATTACHED

8 5

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AN ASSESSMENT OF THE QUALITY OF SAMPLING PROCEDURES REPORTED IN CLINICAL NURSING RESEARCH: A PILOT STUDY

by

Nancy O'Pry Gentry

A paper submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Public Health in the Curriculum in Public Health Nursing.

Chapel Hill

1988

Approved by:

Maija L Selby
Advisor

Dana Quade
Reader

ABSTRACT

NANCY O'PRY GENTRY. An Assessment of the Quality of Sampling Procedures Reported in Clinical Nursing Research: A Pilot Study (under the direction of MAIJA SELBY).

This pilot study assessed the reliability of an instrument specifically designed to assist in the scientific evaluation of the quality of clinical nursing research. This instrument also was used to identify the major errors in sampling in clinical nursing research in a random sample of articles published in selected clinical nursing journals in 1986. A retrospective, nonexperimental pilot study was conducted for 30 articles using the Research Assessment Form (RAF). Content validity and inter- and intra-reliability were established for all but the inferential statistical section of the RAF. Of the 30 articles reviewed, 96.7% contained a major error in sampling, indicating the need for nursing educators to emphasize sampling procedures in their research classes and for publishers and manuscript reviewers to address sampling considerations in their criteria for selecting articles for publication. The majority of the articles did not provide sufficient information to allow for mathematical calculation of statistical power. Although the statistical power for detecting a large effect was good in the 10 studies in which power was estimated, the power for detecting a medium or small effect was low. This suggests the need for modification of the RAF to include a question that could assess

Cont

whether the failure to find statistical significance was related to low power. Refinement of the RAF is indicated to decrease the overall length of the RAF to focus on the most pertinent questions; to include more specific instructions to abstractors in the body of the RAF, in order to decrease the time needed for abstractor training; and to develop a scoring system which would allow for quantitative comparisons of research articles. It also is suggested that research be extended to a larger sample and to other journals or areas of nursing research. This will allow nurses to improve the scientific rigor of their studies and to critically evaluate research.

ACKNOWLEDGEMENTS

The author wishes to express her appreciation to Maija Selby for her guidance and support during each stage of the research process and to Dana Quade for providing valuable statistical expertise regarding the content of this paper. A special thank-you goes to Roberta Riportella-Muller for believing this paper would become a reality. But, the greatest appreciation (and the biggest hugs) go to my family, Lee, Erin, and Corey, who endured me and this paper from beginning to end.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

CHAPTER 1

Introduction

Background

A major goal of any profession is to improve the practice of its members so that the services provided to their clients will have the most positive impact. The development of a scientific body of knowledge is crucial to the attainment of this goal, and can be instrumental in fostering commitment and accountability to the profession's clients (Fawcett, 1980; Leininger, 1976; Merritt, 1986; Polit & Hungler, 1987; Ventura & Waligora-Serafin, 1981).

Research plays a key role in generating the knowledge which provides the foundation or scientific base upon which nursing education and practice are built (American Nurses Association, 1984, 1985; Fitzpatrick & Abraham, 1987; Polit & Hungler, 1987). The establishment of a scientific base of nursing knowledge permits nurses to make more informed decisions in their practice and provides scientific accountability (Polit & Hungler, 1987; Ventura & Waligora-Serafin, 1981). As early as 1948, Dr. Esther Lucile Brown discussed formal research activities and identified the necessity for accountability to clients (Brown, 1948). Professional accountability demands that nurses use the findings of research as a basis for performing their roles and making decisions

(American Nurses Association, 1985; National League for Nursing, 1983,1986; Polit & Hungler, 1987).

According to Fawcett (1980), scientific research is characterized by a clear theoretical base, a systematic and controlled study of concepts and their connections, as well as a critical appraisal of study design and findings (p. 37). An integral feature of study design is the sampling procedure. This procedure is critical in determining the extent to which the findings can be generalized; generalizability is a critical factor in nursing research (Gortner, 1983). The sampling procedure also is important in determining the level of confidence that can be placed in research conclusions. Thus, the sampling procedure affects the capacity to use the findings to augment and change nursing practice (Ganong, 1987; Zalar, 1986).

Other professions have conducted periodic, criterion-based scientific assessment of the methodological quality of their research, including the adequacy of sampling procedures (Brewer, 1972; Brown, Kelen, Moser, Moeschberger, & Rund,1985; Chase & Tucker, 1975; Cohen, 1962; DerSimonian, Charette, McPeek, & Mosteller, 1982; Elenbaas, Cuddy, & Elenbaas, 1983; Emerson, McPeek, & Mosteller, 1984; Fletcher & Fletcher, 1979; Freiman, Chalmers, Smith, & Kuebler, 1978; Gentry & Shulman, 1985; Glantz, 1980; Glass, 1980; Hopkins, 1973; Mosteller, 1979; Rosenthal & Rubin, 1983; Tversky & Kahneman, 1971; Young, Bresnitz, & Strom, 1983). However, even though research

by registered nurses has experienced dramatic growth (Polit & Hungler, 1987; Swanson & McCloskey, 1986) and although numerous critical reviews of nursing research have been published (Abdellah, 1970; Beck, 1985; Brown, Tanner, & Padrick, 1984; Ellis, 1977; Gortner & Nahm, 1977; Highriter, 1977; Hill, Gortner, & Scott, 1980; Jacobsen & Meininger, 1985, 1986; Moustafa, 1985; O'Connell & Duffey, 1978; Schwirian, 1984), these reviews, in general, have not been based on actual scientific assessment of the methodological quality of the research being reviewed. For example, Schwirian (1984) reported that sample sizes in studies of nursing education "usually were quite adequate", but presented no evidence of having assessed the adequacy of the sample sizes through power calculations (p. 224). A major reason for the lack of scientific rigor in the reviews of nursing research has been the lack of a valid and reliable instrument for assessing research quality.

Problem Statement

Therefore, the purpose of this pilot study was to assess the reliability of an instrument specifically designed to assist in the scientific evaluation of the quality of clinical nursing research. A secondary purpose was to use this instrument to identify the major errors in sampling in clinical nursing research in a random sample of articles published in selected clinical nursing journals in 1986.

Justification of Study

According to the American Nurses' Association's Cabinet on Nursing Research (1985) the future of nursing practice and, ultimately, the future of health care in this country depend on nursing research designed to constantly generate an up-to -date, organized body of nursing knowledge. The significance of nursing research will be determined in proportion to the impact that research has on the health needs of the nation (Gortner, Bloch, & Phillips, 1976).

In the United States each year almost 1.5 million registered nurses (USDHHS, 1986) are responsible for a wide variety of health related tasks, e. g., client care, the promotion of health, the prevention of disease, and the mitigation of the effects of acute and chronic illness and disabilities (Merritt, 1986). Nurses provide 24-hour care, client education, and discharge planning for nearly 40 million hospital admissions (American Hospital Association, 1984). They also provide education, assessment, planning, nursing intervention, and evaluation for 273 million hospital outpatient visits (American Hospital Association, 1984); an additional 100,000 registered nurses provide similar services in non-hospital ambulatory settings; 20,000 nurse practitioners and midwives provide primary care (USDHHS, 1986). Another 115,000 registered nurses (USDHHS, 1986) supervise the health care of approximately 1.5 million residents of nursing homes and long-term care facilities (USDHHS, 1984), and 100,000 develop and evaluate health programs in

schools, industries, and community/public health agencies (USDHHS, 1986). If nursing research findings are inappropriately generalized to the populations served by these professional nurses, the potential health effects are immense.

Therefore, because nursing practice does affect all individuals in the United States at some time in their lives, it is imperative to identify and correct any deficiencies in the research upon which nursing practice is based. This pilot study, based upon the standards and principles that govern the research process, is a beginning measure to assess the quality of nursing research. The results of this research are not limited to a single vocation, but can be used by educators, reviewers, editors, and researchers to take specific action to improve sampling techniques in nursing research. Ultimately, this will improve the generalizability of the research findings to nursing practice. The development of a reliable and valid instrument in this study can provide a tool for future use in all areas of nursing research and will allow for the adequate assessment of sampling in future studies.

Literature Review

Fawcett (1980) stated that scientific research is characterized by a clear theoretical base, by a systematic and controlled study of concepts and their connections, and by critical appraisal of study designs and findings (p. 37). The ability to use the findings from research to influence nursing practice is greatly affected by the generalizability of these

findings (Zalar, 1986), and generalizability is greatly influenced by the sample design.

Most researchers who addressed the sample and sampling techniques in their reviews of nursing research did not systematically assess the methodological quality of sampling procedures in the articles they reviewed (Brown, Tanner, & Padrick, 1984; Gortner, 1983; Highriter, 1977; Jacobsen & Meininger, 1985; O'Connell & Duffey, 1978). Other reviewers of nursing research made no reference to the sample or sampling technique (Abdellah, 1970; Ellis, 1977; Gortner & Nahm, 1977; Hill, Gortner, & Scott, 1980; Moustafa, 1985). As mentioned previously, one study (Schwirian, 1984) reported that sample sizes were adequate, yet presented no evidence of having assessed the adequacy of the sample sizes. Jacobsen and Meininger (1986), who conducted the first systematic evaluation of reporting in randomized experiments in nursing, found that only five percent (i. e., two) of 42 reports reviewed provided evidence of sample size or power calculations. Seventy-four percent gave some evidence of pretreatment equivalence of groups, but over half of those claiming equivalence did not provide sufficient evidence to substantiate their claims. Only 50 percent mentioned withdrawals, and only two studies provided sufficient information to determine whether differential withdrawal was a problem (p. 379). As this and other studies reflect, there has been no standardized instrument, to date, developed to objectively or scientifically evaluate the quality of nursing research.

Although Jacobsen and Meininger (1986) provided clear evidence of the need for education of nursing researchers regarding established reporting requirements for clinical trials, they did not address the most disturbing implication of their study: the possibility that the majority of randomized experiments in the most prestigious nursing research journals in the United States contain severe sampling defects. Randomized clinical trials are the "gold standard" against which other research is judged, and the journals reviewed by Jacobsen and Meininger are widely acknowledged to represent the best of nursing research. If nursing's gold standard is imperfect, as Jacobsen and Meininger's findings suggest, then the research foundation upon which nurses are expected to base their clinical practice may be equally flawed (see Appendix A, p. 2).

A systematic evaluation of the quality of sampling in nursing research is required. As pointed out in the landmark study by Jacobsen and Meininger (1986), there is clear evidence of the need to educate nurse researchers about established reporting requirements for clinical trials. Our study builds on the findings of Jacobsen and Meininger (1986) and looks at sampling in more detail. It extends the review beyond clinical trials and examines a variety of research studies in clinical nursing. Furthermore, whereas Jacobsen and Meininger examined research in journals that publish only nursing research and are directed

toward a scientific audience, our study focuses on clinical nursing journals that are more likely to be read by practicing clinicians.

Scientific Rationale

The scientific base for this study is provided by the standards and principles that govern sample selection as described by Abdellah and Levine (1979); Burns and Grove (1987); Cohen (1977); Feinstein (1977); Kerlinger (1986); Kovacs (1985); Pocock (1983); Polit and Hungler (1987); Seaman (1987); Waltz and Bausell (1981); and Wooldridge, Leonard, and Skipper (1978). These standards and principles are summarized below.

It is imperative that the target population and sampling frame be identified. This population comprises the total group of persons or objects that meets the designated set of criteria established by the researcher. And it is to this target population that findings from the study will be generalized. Unless it is quite small, it is impossible to study an entire population. Therefore, research studies typically involve only a small fraction of the population referred to as a sample. The sampling frame is the list of members of the population from which that sample is drawn. Without specific information on the target population and the sampling frame, the reader cannot determine to whom the results of the study are intended to apply.

The sampling method used and a description of either probability or nonprobability sampling should be discussed. It is important to know

the sampling method used in order to determine whether the results can be generalized to the target population.

Probability sampling is a method whereby each person or object in the population has a known chance of being selected for the study. When a probability sample is used, the findings can be generalized from the sample to the population from which the sample was taken. Probability sampling reduces the possibility of selecting a biased sample (i.e., one in which some members of the population are over- or under-represented, and the researcher is not aware of it). The data summaries based on randomly selected samples can be analyzed using statistical techniques that estimate sampling error (i. e., the measure of how much sample findings differ from the true population values). The four most commonly used probability sampling methods are simple random, stratified random, systematic random, and cluster sampling.

Nonprobability sampling must balance the advantages of convenience, economy, and time against the risks involved in not using probability sampling. It is less likely than probability sampling to produce representative samples and accurate estimates because nonprobability samples can be assumed to possess an inherent bias. This does not mean that such studies are bad or that their data are unsatisfactory. It does mean, however, that there is difficulty in generalizing the results from the sample to the population. The major methods used in

nonprobability sampling are accidental or convenience, purposive, and quota sampling.

The response rate and the number of withdrawals and losses (and reasons for these) also should be provided. When the "target" sample and the "actual" sample are not the same, the researcher must not overlook the possibility that the individuals who elect not to participate, or who for some reason cannot participate, would have responded differently from those who do participate (Kovacs, 1985). This information allows the researcher to evaluate the limitations of the sample and whether the results are usable.

A research report must specify the sample size. Every time a researcher calculates a percentage or an average based on sample data, the purpose is to estimate a population value. Smaller samples will tend to produce less precise estimates than larger samples. Thus, in a probability sample, the larger the sample, the more likely it is to be representative of the population (Polit & Hungler, 1987; Selby, 1987). The possibility of achieving statistical significance also is greater with a large sample. However, large samples are by no means an assurance of accuracy; a large sample cannot correct for a faulty sampling design. Furthermore, in most cases, after the sample size increases beyond a certain point, depending both on the way the sample is selected and the characteristic(s) being studied, additional increases produce a useless or unnecessary increase in precision (Remington & Schork, 1985).

Practical restraints such as time, money, and availability of potential subjects also must enter into the sample size decision. The ultimate criterion for assessing a sample is not the quantity of data it produces, but the confidence with which one can make inferences from the sample to a specified population. It is important to state the sample size so that readers can determine for themselves if an inability to find statistical significance is related to an insufficient sample size.

Before a study is undertaken, the sample size must be determined; ideally, such a calculation should be based on statistical measures that ensure the desired levels of three interrelated parameters: power, significance, and effect size. Power is the ability of a research design to detect existing relationships among *variables*, *i. e.*, stating that there is a difference or a relationship when there really is one. The significance level is the probability that an observed relationship could be caused by chance, i. e., because of sampling error. Effect size is a statistical expression of the magnitude of the relationship between two variables, or the magnitude of the difference between two groups, with regard to some attribute of interest (Polit & Hungler, 1987). All of these factors influence sample size.

If the authors present information about sample size, power, significance level, and effect size, the readers can decide how confident they can be about the results of the study. Should the authors not provide information concerning how they calculated the sample size,

other information should be included in the article so that readers can calculate the statistical power of the hypothesis(es) tested. The ability to assess power is particularly important when the null hypothesis is not rejected since failure to reject may be a consequence of low power (in turn, related to small sample size) rather than a true failure of the intervention being tested. This information about power can help researchers decide whether to continue to test an intervention or a relationship (with a larger sample), and can also help clinicians make informed decisions about the possible value of trying an intervention that showed promising but not statistically significant results.

The limitations of the sample, with possible sources and directions of bias, also should be acknowledged. Results of the studies should be generalized appropriately to the sample and/or the sampling frame and not beyond. Stating the limitations will allow the readers to make informed decisions concerning to whom the study results may be applicable.

The RAF incorporates these standards and principles for sampling as presented above. This instrument was used to guide our evaluation of sampling methods and techniques in the articles reviewed. These same standards and principles were used by the researchers in planning the sampling method and selecting the sample for this pilot study.

Assumptions

The assumptions upon which this pilot study of the quality of published clinical nursing research is based are (a) research is the foundation underlying nursing practice (ANA, 1984,1985); (b) the ability to generalize the results of a study from a sample to the target population from which it (i. e., the sample) was drawn is crucial to all research (Abdellah & Levine, 1979; Burns & Grove, 1987; Pocock, 1983; Polit & Hungler, 1987; Waltz & Bausell, 1981; Wooldridge, Leonard, & Skinner, 1978); and (c) published articles which violate scientific principles and established standards for the conduct and reporting of research, specifically in the area of sampling, may jeopardize nursing practice (see Justification, pp. 5-6).

Definition of Terms

In this pilot study, an article with a major error in sampling was defined as one that did not adequately address the following items on the Research Assessment Form (RAF, Appendix B). The responses which constitute inadequate responses are listed under each item below:

1. Error in reporting sampling method: items 29 and 30
   A. Item 29: Sampling method claimed to be used by authors
      response 3 - not addressed

B. Item 30: Sampling method actually described by authors

response 3 - not described sufficiently to categorize

response 4 - not described at all

C. Discrepancy between item 29 (sampling method claimed) and item 30 (sampling method actually described).

2. Error in describing the sample, i. e., sampling frame, sample size, and number of refusals, withdrawals, and/or cases lost: items 31, 32, 33, and 34

A. Item 31: Is sampling frame mentioned (at all, anything)?

response 2 - no

B. Item 32: Total attempted sample size:

response 666666 - unclear or conflicting information

response 999999 - information not provided

C. Item 33: Total completed sample size:

response 666666 or 999999 (see item 32 above)

D. Item 34: Total number of refusals, withdrawals, and/or cases lost:

response 666666 or 999999 (see item 32 above)

3. Error in reporting sample limitations: item 129

   A. Item 129: Were any limitations of sampling stated?

      response 2 - no

4. Error in making generalizations from sample - items 127 and 128

   A. Item 127: Were results generalized beyond the sampling frame?

      response 1 - yes, clearly beyond the sampling frame

      response 2 - unclear

   B. Item 128: Were results generalized beyond the sample? (This was considered to be an error only for articles using nonprobability sampling)

      response 1 - yes

      response 2 - unclear

CHAPTER II

Methodology

Design

A retrospective, nonexperimental pilot study was conducted in which research articles from five clinical nursing journals published during 1986 were reviewed. This particular segment of the study focused specifically on the quality of sampling as reported in these five journals. The study was part of a larger study designed to refine and test the reliability of the RAF and then use the RAF to describe and evaluate the overall quality of nursing research in the selected journals.

Setting

The pilot study was conducted during the academic year 1987-1988 at the School of Public Health, University of North Carolina at Chapel Hill, by faculty and students of the Curriculum in Public Health Nursing and the Department of Biostatistics. Training of data collectors and preliminary testing of the RAF occurred over five sessions during November and December 1987. Five articles were utilized during the training period. Over a period of one month, the data collectors independently completed twenty-five additional articles, using the revised RAF.

Sample

The sampling frame for the study included all 130 original research articles (those which reported data collection and/or analysis) published in 1986 in five United States clinical nursing specialty journals (community health - Public Health Nursing; critical care - Heart & Lung; gerontology - Journal of Gerontological Nursing; maternal and child health - Journal of Obstetric, Gynecologic, and Neonatal Nursing; and oncology - Oncology Nursing Forum). These journals were chosen purposely for three reasons. They provided a variety of clinical specialties; they had a substantial amount of research content (at least 33% of each journal's articles were research studies, as reported by their editors [Swanson & McCloskey, 1986]); and their subscriber circulation (within each specialty) was high.

It was determined that 55 articles were the minimum required to "ensure 95% confidence that the proportion of articles containing a methodological error was within .10 of the true proportion in the sampling frame" (see Appendix A, p. 6, and Appendix D, Part I). To select the 55 articles, the sampling frame of 130 articles was stratified by journal. Then articles were selected proportionately from each stratum, using a table of random digits. From these 55 articles, a random sample of 30 (six from each journal) was selected for the pilot study (see Appendix A, p. 5, for additional specifics concerning the sampling procedure). Thus it was anticipated that results from the pilot study could be generalized to the

sampling frame of 130 research articles published in 1986 in the five selected journals. However, the smaller sample meant that our confidence in the results might be lower than 95% or that our precision would be less than .10 (see Appendix D, Part II).

Instrument

The Research Assessment Form (RAF) was developed by compiling information from a comprehensive literature review (see Appendix C), and was utilized to evaluate the research articles in this pilot study. The RAF offers an objective assessment of each phase of the research process. The areas set forth include: clarity and logic of the purpose statement; use of the literature to justify a need for the study; relevance of the conceptual/theoretical framework to the study; description of the setting; evidence of protection of human subjects; appropriateness of sampling methods and reporting of sample size; adequacy of statistical power; thoroughness of describing the research design (including limitations); documentation of validity and reliability of the instruments utilized; completeness of reporting of descriptive and inferential statistics; thoroughness of reporting the data analysis, results, and implications of the study; and completeness of tables, figures, abstract, and title. The RAF also included information regarding the number of authors, their educational levels, and the type, if any, of grant funding (see Appendix A, p. 6).

Content validity of the RAF was established through a series of reviews by a panel of nationally known experts including a doctorally prepared nurse researcher, a research editor, and a doctorally prepared biostatistician. Reliability was assessed using seven selected articles and seven of the eight sections of the RAF. The seven articles were selected because they offered many potential problems to consider on the RAF. Inter-rater reliability was calculated as a percentage of agreement between the review by the nursing research assistants (RAs) and the "gold standard" review by the principal investigator with biostatistical consultation. The mean percent of agreement with the gold standard was 90.63, standard deviation 1.33; the reliability among raters ranged from 89.5 to 92.1 percent. Intra-rater reliability was established by one of the RAs reviewing five articles (one from each journal), and then reviewing the same five articles again, one week after the first review, without access to the previous review forms.. The intra-rater agreement was 96.1% (Kappa statistic pending). Inter- and intra-rater reliabilities are pending for a portion of the data analysis and results section (Section VI) of the RAF.

Data Collection

The data were collected by three RAs (a second year doctoral biostatistics student and two second year master's public health nursing [PHN] students) under the guidance of the primary investigator (PI) and co-investigator (CI). Training sessions which totaled 12 hours were

performed for initial reliability testing (as described on pp. 19-20) and included three second year master's PHN students and the two investigators.

A subsample of five selected articles, one from each journal, was utilized during the training sessions. Feedback was provided by the PI to the team on the methodological areas questioned by the RAs and/or CI; these discussions among the PI, CI, and RAs resulted in revisions to the RAF.

An additional 25 articles were reviewed by a team of two RAs and the CI working independently of one another. This resulted in a total of 30 articles from the sample, with five of the 30 articles used for training purposes among the three RAs, CI, and PI. The remaining 25 articles were divided among the three reviewers in such a manner that two of the 25 articles were reviewed by all of the investigators; the RAs and CI were not informed of which two articles were used for this quality control check. Inter-rater (gold standard) reliability was calculated for these two articles and they were included in the overall reliability measure (see p. 20). The CI examined all finished RAFs for completeness. Data collection for the remaining 25 articles (of the 55 articles in the larger study) is still in progress.

Data Analysis

A microcomputer was used to analyze the data with SPSS-PC statistical software. Descriptive statistics (means and standard

deviations, ranges, percents, and raw frequencies) were calculated for the variables of interest in the pilot study. The small sample in this pilot study precluded testing for significance of differences, e. g. between journals. In this report the articles first are described according to journal and specialty, number of authors, educational levels and professions of authors, as well as type of funding. Then, the sample is described in terms of the proportion of articles with errors in the area of sampling. Finally, additional information is presented regarding statistical power for those studies that used statistical testing.

Protection of Human Subjects

The authors whose research was examined in this study were considered to have given permission for their studies to be reviewed because public inspection is an expectation of publication. However, in order to avoid any embarrassment to the authors, the data are not reported at the individual author level. The study received Institutional Review Board Approval through the School of Public Health at the University of North Carolina at Chapel Hill (see Appendix E).

CHAPTER III

Results and Conclusions

Results

The RAF was used successfully to review 30 randomly selected articles (six articles from each of the five selected specialty journals). The mean time for completion of the RAF for each article was 54 minutes (SD 24 minutes; median = 45 minutes, with a range from 25 to 101 minutes). After training sessions, consultation with the PI and statistical consultant was required for five (20%) articles.

Characteristics of Authors

Table 1 describes the characteristics of the authors for the articles reviewed. Nearly all first authors were registered nurses. The master's degree was the highest degree held by approximately two-thirds of the first authors. Approximately one third of the articles were written by a single author. The maximum number of authors was five.

Most (73.3%) of the research was not grant funded. Of the eight articles which received grant funding, 1 (3.3% of the total articles) received federal funding, 5 (16.7% of the total articles) received non-federal funding, and 2 (6.7% of the total articles) received both federal and nonfederal funding.

Sample

Error in reporting sampling method

A large majority of the authors claimed to use and, according to descriptions in the articles, actually used nonprobability sampling (Table 2). Four of the articles did not address the sampling method used. It was evident that two of these four actually used nonprobability sampling. All three studies reporting probability sampling did provide evidence that they actually used probability sampling.

Error in describing sampling frame, sample size, and number of refusals, withdrawals, and/or cases lost

The authors in 25 (83.3%) of the 30 articles provided some information about the sampling frame in their studies (Table 3). However, in 5 (16.7%) of the articles, it was impossible to discern any information about the population from which the sample was selected.

Of the 30 articles addressing sample size (Table 3), only 1 (3.3%) of the studies provided any evidence of having used mathematical calculations for determining sample size. Of the 16 studies whose main purpose was to describe a population, none provided estimates of the population proportion, population size, desired level of confidence, or desired level of precision. Of the 14 studies whose main purpose was to describe differences between groups, none provided evidence of using the significance level or power to calculate sample size. However, one

study provided evidence of using effect size; thus it could be inferred that significance and power were considered, though not reported.

Two-thirds (n=20) of the authors did not provide information on the attempted sample size in their studies (Table 4); 2 provided unclear information regarding the completed sample size (Table 5). Also, the number of refusals, withdrawals, or cases lost was either unclear or not provided in 18 (60%) of the articles (Table 6).

Errors in reporting sample limitations

Sampling limitations were not stated in 13 (43.3%) of the articles (Table 7). Table 8 provides a listing of the types of sampling limitations which were evident to the reviewers, but not acknowledged by the authors of the studies.

Error in making generalizations from sample

Tables 9 and 10 provide information regarding the generalizations made by the authors in the articles reviewed. Over half of the 27 studies using nonprobability sampling inappropriately generalized their results beyond their sample. Furthermore, 30% (n=9) of all of the studies included inappropriate generalization of results beyond the sampling frame.

Summary of errors in sampling

Overall, 29 (96.7%) of the 30 articles reviewed contained at least one of the four major errors in sampling (Table 11). Over two-thirds of the articles had an error in describing sampling frame, sample size, or

number of refusals, withdrawals, and/or cases lost; more than one-half committed an error in making generalizations from the sample.

Power of statistical tests in samples used

Nineteen of the 30 articles reported statistical testing; the remaining 11 were descriptive only and therefore did not report any statistical tests. Four of the 19 articles had insufficient information for assessing statistical power; the data collector was unable to assess statistical power in 2 other articles which currently are being assessed with biostatistical consultation. Three of the remaining 13 articles provided sufficient information to allow mathematical calculation of power. For the remaining 10 articles, power was estimated using Cohen's (1977) tables for a small, medium, and large effect. Table 12 shows that the specific types of information needed for mathematical calculations of power generally were not provided in the articles assessed.

For the three articles in which exact power calculation was possible, the power for statistical tests for the first major hypothesis in each article was .10, .20, and .82, respectively. Table 13 provides the power estimates for small, medium, and large effect for the 10 articles for which power was estimated based on Cohen's (1977) tables. For a small effect size, all power estimations for the first major null hypothesis were <.40; mean power was extremely low ($\overline{x}$ = .16 ± SD .20). For a medium effect, half of the studies had statistical power less than .50 ($\overline{x}$ = .58 ± SD

.32). However, for a large effect, only 2 (20%) of the 10 studies had power of less than .5; 7 (70%) had power above .7, and 6 (60%) had power of .9 or above. Mean power for large effect size was high ($\bar{x}$ = .80 ± SD .88), but the range was large (Table 13).

For the 5 articles reporting testing of a second major hypothesis, all power estimations for a small effect size were <.30 ($\bar{x}$ = 0.11 ± SD .05). For a medium effect, power estimates ranged from .18 to .99. For large effect, 4 of the 5 studies had statistical power above .80 ($\bar{x}$ =.81 ± SD .75) (Table 14).

In 7 (36.8%) of the 19 studies, the authors did not provide sufficient information to determine whether their decisions regarding rejection of the null hypothesis was appropriate. In 4 (21.7%) of the 19 articles, the author(s) stated that failure to reject the null hypothesis proved that the null hypothesis was true (Table 15). In 1 of these, the actual power for detecting statistical significance of the observed effect was only .10. Details regarding the sample size and power in these 4 studies are provided in Table 16. The RAF did not assess the total number of studies that failed to reject a null hypothesis; it asked only whether the authors put great emphasis on the importance of "accepting" their null hypothesis.

Limitations

This study itself is subject to human error since the data must be extrapolated from the research articles. Several control measures were created to minimize the possibility of errors. These measures included

training sessions, reliability testing, procedural quality controls, and expert consultation. However, even with these control measures, there were still some areas of the RAF which required additional consultation to correct reviewer error or questions. One portion of data collection (regarding statistical power) has not yet been completed, so that presently all the data are not available for analysis.

Initially it was believed that the pilot study sample size was a major limitation and that the results of the pilot study could be generalized with confidence only to the sample. In planning the study of 55 articles, it was estimated that we could be 95% confident that the results regarding the proportion of studies with 1 or more major methodological errors would be within ± .10 of the true population proportion (based on an estimated probability that 50% of the articles would have a major error). Using a sample of only 30, our anticipated level of confidence, given an estimated proportion of 50% and desired precision of ±.10, was less than 80% (see Appendix D). If we wanted to continue to have 95% confidence, our precision was less exact, nearly ±.16 (see Appendix D). However, our actual estimated proportion of articles with a major error in sampling was 96.7%, not 50%. This resulted in a more precise estimate than expected; in out study, we can be 95% confident that the proportion of studies we found with errors in sampling methodology, 96.7% in our sample, might actually be 97% ± 5.6%(91%, 100%) in the population of 130 research articles (Appendix F). Nevertheless, we cannot generalize to any year

other than 1986 in the selected journals, nor to any journals other than those studied.

Discussion and Implications

This pilot study provided information to evaluate the quality of nursing research published in 1986 in five major clinical nursing specialty journals, in terms of adherence to the standards and principles that govern sample selection. Although the sample we reviewed was itself quite small, we did use a probability sample and found patterns that suggested serious problems in relation to sampling methodology. Almost all of the articles reviewed contained at least one of the four major errors in sampling. These errors could severely jeopardize the validity of the results in the articles reviewed and could cause unfortunate repercussions if the results from these studies are inappropriately used in nursing practice.

A majority of the articles used nonprobability sampling, which severely limits the generalizability of the findings. Nonprobability samples are likely to be more biased and produce less representative and less accurate results than probability samples (Gentry and Shulman, 1985). Unfortunately, in our sample, over half of the articles using nonprobability sampling generalized their results beyond the sample; nearly 30% of all articles even generalized their results beyond the sampling frame. Furthermore, sampling limitations were not stated in over a third of the articles. It is difficult for the reader to ascertain the

generalizability of results in studies in which the authors neglect to discuss limitations. From these findings it appears that many researchers did not realize that nonprobability sampling precludes generalization beyond the sample, and that even probability sampling will not allow generalization beyond the sampling frame. A number of authors also did not appear to know that the specific demographic characteristics of their sample subjects could confound their results. Therefore, we would suggest that nursing educators emphasize these points in their research classes. It would also be helpful if journal editors and manuscript reviewers were to incorporate the standards and principles that govern the sampling plan into their criteria for selecting articles for publication.

Although the sampling frame was identified in a majority of the articles, most of these articles did not communicate the attempted sample size, the rationale for the sample size, or the sample size calculations. Consequently, readers are unable to determine the target population for whom the study results were intended to apply or if an inability to find statistical significance was related to an insufficient sample size. Furthermore, almost all the articles omitted data that would have allowed for mathematical calculation of the power for detecting statistical significance of the results in the given sample. This included omitting sample size, the name of the statistical test, the alpha level, and/or the effect size. Although full disclosure of all information regarding sample size and power calculations is rare in the literature, our findings

suggested that the authors may not even have realized that a priori sample size and power calculations are desirable (or even possible). This finding is reinforced by our own discussions with other students and faculty in schools of nursing, who report that sample size calculation usually is not taught. Most nursing research texts omit or "gloss over" the topic. Therefore, we would suggest that procedures for calculating sample size and guidelines for conferring with statisticians about sample size be included in nursing research courses. Journal editors and manuscript reviewers also could help by addressing sample size considerations in their criteria for selecting articles for publication.

In our study, although the authors of most of the studies we reviewed did not provide sufficient information for us to perform mathematical calculation of statistical power, for 10 of the studies we were able to estimate power for small, medium, and large effect from the tables prepared by Cohen (1977). Although the statistical power for detecting a large effect was good ($\overline{x} > .80$) in our sample, the power for detecting a medium or small effect was low. Regrettably, in this study we did not determine the overall number of studies that failed to reject a null hypothesis. If we had done this, we could ave assessed whether the failure to find statistical significance was related to low power. Therefore, it is suggested that the RAF be modified to include a question relating to rejecting or failing to reject a null hypothesis, ana that research be undertaken to assess this issue in our sample as well as in other

samples. Furthermore, because the sample for which power was calculated was extremely small, we would advise extending our research to a larger sample.

We found that the RAF does have content validity. We also established excellent inter- and intra-rater reliability for all but one section of the RAF; continued testing is needed to complete the reliability assessment for the inferential statistical portion of the RAF. Upon completion of the reliability assessment, the RAF can be utilized by nurse researchers and nursing students in order to critically review research studies. In addition, it has the potential for use by reviewers and publishers as a guide or checklist of criteria which must be incorporated into research studies for publication. Additional study is indicated to refine the instrument as follows: (a) to decrease the overall length of the RAF to include only the most pertinent questions; (b) to include more specific instructions to abstractors in the body of the RAF, in order to decrease the time needed for abstractor training; and (c) to develop a scoring system which would allow for quantitative comparisons of research articles. We also would suggest further research to extend this tool to other journals or areas of nursing research. The pilot study is in the process of being extended to incorporate the remaining 25 articles from the original calculated sample size of 55 articles.

Future research will enable the profession to monitor itself regarding the quality of research and identify those areas which require

further refinement. The data base created through this project also will provide a methodologically sound foundation for future statistical evaluation of changes in research quality over time (see Appendix A, p. 3).

Conclusions

As with any research, one must be cautious in interpreting and generalizing the results of this study. The results of this pilot study can be generalized only to the 130 research articles published in 1986 in the five clinical nursing specialty journals selected for study; however, the general consistency of the findings suggest that there may be serious methodological problems with sampling in clinical nursing research. This pilot study provided a *beginning evaluation of the research* that is likely to be read by practicing clinicians and established a methodologically sound foundation for future research (see Appendix A, p. 8).

Since research is an expectation of practice, and nurses are expected to base their practice on research, the potential significance of basing practice on flawed research is tremendous. This pilot study has revealed a substantial number of the methodological errors in sampling and has provided suggestions for changes in education , research, and publishing practices. This will allow nurses to improve the scientific rigor of their studies and to critically evaluate research. Ultimately, practitioners and their clients will benefit from the findings of improved research.

Table 1

Selected Characteristics of Authors in 30 Research Articles Reviewed

	Articles (n=30)	
Author characteristics	n	%
Profession of first author		
RN	28	93.3
Non-RN	1	3.3
Information not provided	1	3.3
Highest degree of first author		
Doctorate	10	33.3
Master's	18	60.0
Information not provided	2	6.7
Number of authors		
1	9	30.0
2	10	33.3
3 or 4	10	33.3
5	1	3.3

Table 2

Sampling Methods Claimed to be Used and Sampling Methods Actually Used in 30 Research Articles Reviewed

	Articles (n=30)	
Sampling method	n	%
Claimed by author(s)		
Probability	3	10.0
Nonprobability	23	76.7
Not addressed	4	13.3
Actually used[a]		
Probability	3	10.0
Nonprobability	25	83.3
Not addressed	2	6.7

Note. [a]as determined from information presented in article.

Table 3

Reporting of Information Regarding Sampling Frame and Rationale for Sample Size Selection in 30 Research Articles Reviewed

	Articles (n=30)	
	n	%
Identification of sampling frame		
At least mentioned in article	25	83.3
Not identified at all	5	16.7
Reporting of rationale for sample size selection		
No rationale or "practical" rationale	29	96.7
Reported evidence of sample size calculations	1	3.3

Table 4

Reporting of Information Regarding Attempted Sample Size in 30 Research Articles Reviewed

	Articles (n=30)	
	n	%
Reporting of attempted sample size		
Reported = 1-100[a]	4	13.3
Reported =101-500[b]	5	16.7
Reported >500[c]	1	3.3
Information unclear/not provided	20	66.7

Note. [a]sample size = 13, 32, 48, 64

[b]sample size = 141, 296, 300, 337, 343

[c]sample size = 1689

Table 5

Reporting of Information Regarding Completed Sample Size in 30 Research Articles Reviewed

	Articles (n=30)	
	n	%
Reporting of completed sample size		
Reported = 1-100[a]	18	60.0
Reported = 101-500[b]	9	30.0
Reported >500[c]	1	3.3
Information unclear	2	6.7

Note: [a]sample size = 12(x2), 14, 20, 25, 30, 31, 40, 45(x2), 48, 50, 58, 60, 62, 74, 80

[b]sample size = 100, 112, 125, 152, 197, 202, 211, 213, 217, 275

[c]sample size = 1250

Table 6

Reporting of Information Regarding Number of Refusals, Withdrawals, and/or Cases Lost in 30 Research Articles Reviewed

	Articles (n=30)	
	n	%
Reporting of refusals, withdrawals, &/or cases lost		
Reported = 1-100[a]	9	30.0
Reported = 101-439[b]	3	10.0
Information unclear/not provided	18	60.0

Note. [a]number reported = 1(x2), 3 (x2), 4, 20, 29, 62, 98

[b]number reported = 146, 246, 439

Table 7

Reporting of Sampling Limitations in 30 Research Articles Reviewed

	Articles (n=30)	
	n	%
Author(s) stated at least 1 limitation		
Yes	17	56.7
No	13	43.3

Table 8

Types of Sampling Limitations Not Addressed by Author(s) in 30 Research Articles Reviewed

Did not acknowledge	Articles n[a]
Limitations of small setting	3
Limitations of convenience sample	8
Confounding factors associated with demographics	8
Number of refusals, withdrawals, and/or cases lost	2
Influence of the Hawthorne effect on subjects	1
Limitations of sample size	7

Note. [a]$n \neq 30$; some articles had more than one limitation.

Table 9

Authors' Generalizations of Results in 30 Research Articles Reviewed

	Articles (n=30) n	%
Author(s) generalized beyond sample		
Yes	17	56.7
Unclear	1	3.3
No	12	40.0
Author(s) generalized beyond sampling frame		
Yes	9	30.0
Unclear	2	6.7
No	19	63.3

Table 10

Appropriateness of Authors' Generalizations of Results in 30 Research Articles Reviewed, Based on Sampling Method Used

	Articles					
	Total (n=30)		Probability Sample (n=3)		Nonprobability Sample (n=27)	
	n	%	n	%	n	%
Were findings generalized						
Beyond sample?	17	56.7	3	100.0	14	51.9[a]
Beyond sampling frame?	9	30.0	1	33.3[a]	8	29.6[a]

Note. [a]inappropriate generalization based on sampling method

Table 11

Summary of Major Errors in Sampling Procedures in 30 Research Articles Reviewed[a]

	Articles (n=30)	
Type of major error in a sampling procedure	n[b]	%[b]
Error in reporting sampling methods	4	13.3
Error in describing sampling frame, sample size, and number of refusals, withdrawals, and/or cases lost	21	70.0
Error in reporting sampling limitations	12	40.0
Error in making generalizations from sample	16	53.3

Note. [a]Overall, 29 (96.7%) of the 30 articles reviewed contained at least one of the four types of errors.

[b]$n \neq 30$; some articles had more than one major error.

Table 12

Reporting of Information Required for Mathematical Calculations of Statistical Power for 13[a] Articles Reviewed

	Articles (n=13)	
Information needed	n	%
Sample size		
Missing	1	7.7
Not missing	12	92.3
Name of statistical test		
Missing	2	15.4
Not missing	11	84.6
Alpha level		
Missing	2	15.4
Not missing	11	84.6
Effect size		
Missing	13	100.0
Not missing	0	0.0

Note. [a]19 articles reported statistical testing; however, in this study, power could be assessed for only 13 of these

Table 13

Statistical Power for Small, Medium, and Large Effect[a] for First Major Hypothesis in 10[b] Research Articles Reporting at Least 1 Statistical Test

	Effect					
	Small[c]		Medium[d]		Large[e]	
Power	n	%	n	%	n	%
<.10	4	40.0	0	0.0	0	0.0
.10 - .19	3	30.0	1	10.0	0	0.0
.20 - .29	0	0.0	2	20.0	0	0.0
.30 - .39	3	30.0	0	0.0	1	10.0
.40 - .49	0	0.0	2	20.0	1	10.0
.50 - .59	0	0.0	0	0.0	1	10.0
.60 - .69	0	0.0	1	10.0	0	0.0
.70 - .79	0	0.0	1	10.0	1	10.0
.80 - .89	0	0.0	0	0.0	0	0.0
.90 - .99	0	0.0	3	30.0	6	60.0

Table 13 continued on next page

Table 13 (continued)

Note. [a]Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

[b]19 articles reported statistical testing; however, in this study, power could be assessed for only 13. Mathematical power calculations were performed for 3 of the 13 (see p. 26)

[c]$\bar{x} \pm sd = .16 \pm .20$

[d]$\bar{x} \pm sd = .58 \pm .32$

[e]$\bar{x} \pm sd = .80 \pm .88$

Table 14

Statistical Power for Small, Medium, and Large Effect[a] for Second Major Hypothesis in 5[b] Research Articles Reporting at Least 2 Statistical Tests

	Effect					
	Small[c]		Medium[d]		Large[e]	
Power	n	%	n	%	n	%
<.10	1	20.0	0	0.0	0	0.0
.10 - .19	3	60.0	1	20.0	0	0.0
.20 - .29	1	20.0	0	0.0	0	0.0
.30 - .39	0	0.0	0	0.0	1	20.0
.40 - .49	0	0.0	2	40.0	0	0.0
.50 - .59	0	0.0	0	0.0	0	0.0
.60 - .69	0	0.0	0	0.0	0	0.0
.70 - .79	0	0.0	1	20.0	0	0.0
.80 - .89	0	0.0	0	0.0	2	20.0
.90 - .99	0	0.0	1	20.0	2	20.0

Table 14 continued on next page

Table 14 (continued)

Note. [a]Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

[b]only 5 of the 10 articles shown in Table 13 reported >1 statistical test

[c]$\overline{x} \pm sd = .11 \pm .05$

[d]$\overline{x} \pm sd = .55 \pm .62$

[e]$\overline{x} \pm sd = .81 \pm .75$

Table 15

Decisions Regarding Testing of Null Hypothesis in 19[a] Research Articles Reporting Statistical Testing

	Articles (n=19)	
	n	%
Decision regarding rejection of null hypothesis		
Correct decision	12	63.2
Insufficient information provided	7	36.8
Author stated that failure to reject the null hypothesis proves the null hypothesis is true		
Yes	4	21.1
No	15	78.9

Note. [a]Eleven of the 30 articles had no statistical testing; therefore, n=19.

Table 16

Sample Sizes and Statistical Power for 4 Articles in Which Authors Emphasized That Failure to Reject the Null Hypothesis Proved That the Null Hypothesis Was True

		Power			
			Estimates[a]		
Article	Sample Size	Mathematical Calculation	For Small Effect	For Medium Effect	For Large Effect
1	20	N/A	N/A	N/A	N/A
2	40	.10	--	--	--
3	62	--	.12	.66	.99
4	211	--	.30	.99	.99

Note. [a]Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

N/A - not available at this time; currently being assessed with biostatistical consultation

-- If exact power was calculable mathematically, this was done. Otherwise, power estimates were made for small, medium, and large effects.

REFERENCES

Abdellah, F. G. (1970). Overview of nursing research 1955-1968, part 1. Nursing Research, 19, 6-17.

Abdellah, F. G., & Levine, E. (1979). Better patient care through nursing research (2nd ed.). New York: MacMillan Publishing Co., Inc.

American Hospital Association (1984). Hospital statistics (1984 ed.). Chicago: Author.

American Nurses' Association (1984). Issues in professional nursing practice: 3. Theory and research as basic to nursing practice. Kansas City, MO: Author.

American Nurses' Association (1985). Directions for nursing research: Toward the twenty-first century. Kansas City, MO: Author.

Beck, C. T. (1985). Theoretical frameworks cited in Nursing Research from January 1974-June 1985. Nurse Educator, 10(6), 36-38.

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 9, 391-401.

Brown, C. G., Kelen, G. D., Moser, M., Moeschberger, M. L., & Rund, D. A. (1985). Methodology reporting in three acute care journals: Replication and reliability. Annals of Emergency Medicine, 14, 986-991.

Brown, E. L. (1948). Nursing for the future. New York: Russell Sage Foundation.

Brown, J. S., Tanner, C. A., & Padrick, K. P. (1984). Nursing's search for scientific knowledge. Nursing Research, 33, 26-32.

Burns, N., & Grove, S. K. (1987). The practice of nursing research. Conduct, critique and utilization. Philadelphia: W. B. Saunders Company.

Chase, L. J., & Tucker, R. K. (1975). A power-analytic examination of contemporary communication research. Speech Monographs, 42(3), 29-41.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

DerSimonian, R., Charette, L. J., McPeek, B., & Mosteller, F. (1982). Reporting on methods in clinical trials. The New England Journal of Medicine, 306, 1332-1337.

Elenbaas, J. K., Cuddy, P. G., & Elenbaas, R. M. (1983). Evaluating the medical literature, part III: Results and discussion. Annals of Emergency Medicine, 12, 679-686.

Ellis, R. (1977). Fallibilities, fragments, and frames: Contemplation on 25 years of research in medical-surgical nursing. Nursing Research, 26, 177-182.

Emerson, J. D., McPeek, B., & Mosteller, F. (1984). Reporting clinical trials in general surgical journals. Surgery, 95, 572-579.

Fawcett, J. (1980). A declaration of nursing independence: The relation of theory and research to nursing practice. Journal of Nursing Administration, 10(6), 36-39.

Feinstein, A. R. (1977). Clinical biostatistics. St. Louis: C. V. Mosby.

Fitzpatrick, J. J., & Abraham, I. L. (1987). Toward the socialization of scholars and scientists. Nurse Educator, 12(3), 23-25.

Fletcher, R. H., & Fletcher, S. W. (1979). Clinical research in general medical journals. A 30-year perspective. The New England Journal of Medicine, 301, 180-183.

Freiman, J. A., Chalmers, T. D., Smith, Jr., H., & Kuebler, R. R. (1978). The importance or beta, the type II error and sample size in the design and interpretation of the randomized control trial. The New England Journal of Medicine, 299, 690-694.

Ganong, L. H. (1987). Integrative reviews of nursing research. Research in Nursing & Health, 10, 1-11.

Glantz, S. A. (1980). Biostatistics: How to detect, correct and prevent errors in the medical literature. Circulation, 61, 1-7.

Glass, G. V. (1980). Summarizing effect sizes. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 13-32). San Francisco: Jossey-Bass Inc., Publishers.

Gortner, S. R. (1983). The history and philosophy of nursing science and research. Advances in Nursing Science, 5(2), 1-8.

Gortner, S. R., & Nahm, H. (1977). An overview of nursing research in the United States. Nursing Research, 26, 10-33.

Highriter, M. E. (1977). The status of community health nursing research. Nursing Research, 26, 183-192.

Hill, M. S., Gortner, S. R., & Scott, J. M. (1980). Educational research in nursing - an overview. International Nursing Review, 27, 10-17.

Hopkins, K. D. (1973). Preventing the number one misinterpretation of behavioral research, or how to increase statistical power. The Journal of Special Education, 7(1), 103-107.

Jacobsen, B. S., & Meininger, J. C. (1985). The designs and methods of published nursing research: 1956-1983. Nursing Research, 34, 306-312.

Jacobsen, B. S., & Meininger, J. C. (1986). Randomized experiments in nursing: The quality of reporting. Nursing Research, 35, 379-382.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart, and Winston.

Kovacs, A. R. (1985). The research process: Essentials of skill development. Philadelphia: F. A. Davis Company.

Leininger, M. (1976). Doctoral programs for nurses: Trends, questions, and projected plans. Nursing Research, 25, 201-210.

Merritt, D. H. (1986). The National Center for Nursing Research. Image: Journal of Nursing Scholarship, 18, 84-85.

Mosteller, F. (1979). Problems of omission in communications. Clinical Pharmocology and Therapeutics, 25, 761-764.

Moustafa, N. G. (1985). Nursing research from 1977-1981. Western Journal of Nursing Research, 7, 349-356.

National League for Nursing (1983). Criteria for the appraisal of baccalaureate and higher degree programs (5th ed.). New York: Author.

National League for Nursing (1986). Characteristics of master's education in nursing (3rd revision). New York: Author.

O'Connell, K. A., & Duffey, M. (1978). Research in nursing practice: Its present scope. In N. L. Chaska (Ed.), The nursing profession (pp. 161-174). New York: McGraw Hill.

Pocock, S. J. (1983). Clinical trials: A practical approach. New York: John Wiley & Sons.

Polit, D. F., & Hungler, B. P. (1987). Nursing research: Principles and methods (3rd ed.). Philadelphia: J. B. Lippincott Company.

Remington, R. D., & Schork, M. A. (1985). Statistics with applications to the biological and health sciences (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rosenthal, R., & Rubin, D. B. (1983). Comparing effect sizes of independent studies. In R. J. Light (Ed.). Evaluation studies review annual (pp. 235-239). Beverly Hills: Sage Publications.

Schwirian, P. M. (1984). Research on nursing students. In H. H. Werley, J. J. Fitzpatrick, & R. L. Taunton (Eds.). Annual review of nursing research (pp. 211-237). New York: Springer Publishing Company.

Seaman, C. H. C. (1987). Research methods. Principles, practice, and theory for nursing (3rd ed.). Norwalk, CT: Appleton & Lange.

Selby, M. L. (1987). Research study guide. Chapel Hill, NC: Curriculum in Public Health Nursing, University of North Carolina at Chapel Hill.

Swanson, E., & McCloskey, J. (1986). Publishing opportunities for nurses. A comparison of publishing policies and practices of 139 journals. Nursing Outlook, 34, 227-235.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

U. S. Department of Health & Human Services (1984). Trends in nursing and related care homes and hospitals. National Health Survey, Series 14, No. 30. Hyattsville, MD: U. S. Government Printing Office, No.(PHS)84-1825.

U. S. Department of Health & Human Services (1986). The registered nurse population. National sample survey of registered nurses, November, 1984. Springfield, VA: National Technical Information Center, Accession No. HRP-0906938.

Ventura, M. R., & Waligora-Serafin, B. (1981). Setting priorities for nursing research. The Journal of Nursing Administration, 11(6), 30-34.

Waltz, C., & Bausell, R. B. (1981). Nursing research: Design, statistics and computer analysis. Philadelphia: F. A. Davis Company.

Wooldridge, P. J., Leonard R. C., & Skipper, J. K. (1978). Methods of clinical experimentation to improve patient care. St. Louis: C. V. Mosby Company.

Young, M. J., Bresnitz, E. A., & Strom, B. L. (1983). Sample size *nomograms for interpreting negative clinical studies.* Annals of Internal Medicine, 99, 248-251.

Zalar, M. K. (1986). Research: Why knowledge about sampling is important. Perioperative Nursing Quarterly, 2, 74-79.

BIBLIOGRAPHY

American Psychological Association (1984). Publication manual of the American Psychological Association (3rd ed.). Washington, D. C.: Author.

Bergman, R. (1984). Omissions in nursing research. International Nursing Review, 31, 55-56.

Bond, S., & Bond, J. (1982). A delphi survey of clinical nursing research priorities. Journal of Advanced Nursing, 7, 565-575.

Chalmers, T. C., Smith, Jr., H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. Controlled Clinical Trials, 2, 31-49.

Cobb, E. B. (1985). Planning research studies: An alternative to power analysis. Nursing Research, 34, 386-388.

Downs, F. S., & Fleming, J. W. (1979). Issues in nursing research. New York: Appleton-Century-Crofts.

Fagin, C. M. (1982a). The economic value of nursing research. American Journal of Nursing, 82,1844-1849.

Fagin, C. M. (1982b). The quality of nursing journals as rated by deans of nursing schools. Heart & Lung, 11, 65-68.

Farrell, M., & Bhaduri, A. (1977). The fundamentals of sampling techniques. Nursing Journal of India, 68, 278-279.

Fawcett, J. (1979). Integrating research into the faculty workload. Nursing Outlook, 27, 259-262.

Feldman, H. R. (1980). Nursing research in the 1980's: Issues and implications. Advances in Nursing Science, 3, 85-92.

Felson, D. T., Cupples, L. A., & Meenan, R. F. (1984). Misuse of statistical methods in Arthritis and Rheumatism. Arthritis and Rheumatism, 27, 1018-1022.

Ford, L. C. (1979). A nurse for all settings: The nurse practitioner. Nursing Outlook, 27, 516-521.

Gentry, M., & Shulman, A. D. (1985). Survey of sampling techniques in widowhood research, 1973-1983. Journal of Gerontology, 40, 641-643.

Goodwin, L. D. (1984). The use of power estimation in nursing research. Nursing Research, 33, 118-120.

Gortner, S. R. (1980). Nursing research: Out of the past and into the future. Nursing Research, 29, 204-207.

Gortner, S. R., Bloch, D., & Phillips, T. R. (1976). Contributions of nursing research to patient care. Journal of Nursing Administration, 6(3), 22-28.

Gothler, A. M. (1986). Quality nursing research: The key to quality nursing education. Nurse Educator, 11(2), 14-16.

Harrell, J. S. (1986). Needed: Nurse engineers to link theory and practice. Nursing Outlook, 34, 196-198.

Hunt, M. (1987). The process of translating research findings into nursing practice. Journal of Advanced Nursing, 12, 101-110.

Issac, S., & Michael, W. B. (1985). Handbook in research and evaluation (2nd ed.). San Diego: EdITS publishers.

Kaiser, D. L., & Veney, J. E. (1980). Describing the sample data. Infection Control, 1, 185-188.

Larson, E. (1984). The current status of nursing research. Nursing Forum, 11, 131-134.

Rittenmeyer, P. A. (1982). The evolution of nursing research. Western Journal of Nursing Research, 4, 223-225.

Sachs, B. (1987). Sample selections in nursing research. Research in Nursing & Health, 10, iii-iv.

Smith, M. C., & Horns, P. N. (1987). Forces guiding the future of nursing research. Nursing & Health Care, 8, 23-25.

Soeken, K. L. (1985). Critiquing research. Steps for complete evaluation of an article. AORN Journal, 41, 882-893.

Tornquist, E. M. (1986). From proposal to publication. An informal guide to writing about nursing research. Menlo Park, CA: Addison-Wesley Publishing Company.

Appendix A

Selected Excerpt from Sigma Theta Tau Grant Proposal

Testing a Tool to Assess Research Quality in Nursing

A. Specific Aims

The major aim of this project is to refine and test the reliability of the Research Assessment Form (RAF), an instrument designed to identify methodological and statistical errors in published nursing research. An additional aim is to use the RAF to describe and evaluate the quality of research published in five major clinical nursing journals in 1986. This is a first step in establishing an ongoing program to monitor the quality of the research upon which nurses are expected to base their practice.

B. Significance

Nurses are increasingly encouraged and expected to use research as the basis for decisions in professional practice (American Nurses' Association, 1985; National League for Nursing, 1983, 1986), and the amount of research published by nurses has increased dramatically (Swanson & McCloskey, 1986). As other professions have embarked on research agendas, they have conducted periodic, criterion-based scientific assessments of the methodological quality of their research (see RAF reference list, Appendix B). Results of such assessments have led to changes in educational and publishing practices directed at improving research quality (O'Fallon et al., 1978; Schor & Karten, 1966). The nursing profession has published numerous critical reviews of nursing research (Abdellah, 1970; Beck, 1985; Benoliel, 1983; Brown, Tanner, & Padrick, 1984; Cronenwett, 1982; Diers & Molde, 1979; Ellis, 1977; Gortner, 1983; Gortner & Nahm, 1977; Grant & Padilla, 1985; Henderson, 1957; Highriter, 1977; Jacobsen & Meininger, 1985, 1986; Lindsey, 1982, 1983; Loomis, 1985; McCloskey, 1981; Moustafa, 1985; O'Connell, 1983; O'Connell & Duffey, 1978; Pollock, 1987; Schwirian, 1984; Stehle, 1981; Werley, Fitzpatrick, & Taunton, 1983, 1984, 1985, 1986). However, for the most part these reviews have not been based on systematic evaluation of the methodological quality of the research being reviewed (Ganong, 1987).

For example, Schwirian (1984) reported that sample sizes in studies of nursing education "usually were quite adequate" (p. 224), but presented no evidence of having assessed the adequacy of the sample sizes through power calculations. Jacobsen and Meininger (1985) described but did not evaluate research methods in nursing research journals from 1956 through 1983. Benoliel (1983) and Brown and colleagues (1984) concluded that research methods had become more sophisticated and sound, based on increases in the use of statistical tests, particularly multivariate analyses; but they did not evaluate whether the statistics were used appropriately. Findings from other professions (Felson, Cupples, & Meenan, 1984) suggest that increased use of statistical tests results in more opportunities for misuse. More recently, Jones and Jones (1987) reported that only 51% of research reports in psychiatric nursing used "correct" statistical procedures; however, their review did not identify statistical errors, nor was it replicable as reported.

In a landmark study, Jacobsen & Meininger (1986) conducted a scientific evaluation of reports of experiments in nursing. Utilizing four established standards for reporting controlled trials (Chalmers et al., 1981; Hill, 1971; Pocock, 1983), these researchers evaluated 42 reports of randomized controlled studies published in three nursing research journals from 1980 through 1984. The research fared poorly on all four criteria: only 5% (i.e., two) of the

reports provided evidence of sample size or power calculation, and only 14% reported the method of random assignment; 74% gave some evidence of pre-treatment equivalence of groups, but over half of those claiming equivalence did not provide sufficient evidence to substantiate their claims; 50% mentioned withdrawals, but only two provided sufficient information to determine whether differential withdrawal was a problem.

Jacobsen and Meininger (1986) provided clear evidence of the need to educate nursing researchers about established reporting requirements for clinical trials, but they did not address the most disturbing implication of their study: the possibility that the majority of randomized experiments in the most prestigious nursing research journals in the United States are seriously flawed. Randomized clinical trials are the gold standard against which other research is judged, and the journals reviewed by Jacobsen and Meininger are the gold standard of nursing research journals. These journals are written for and read by researchers; the research that most clinical nurses read is in clinical specialty journals, not these prestigious research journals. Therefore, if nursing's gold standard is as tarnished as Jacobsen and Meininger's findings suggest, the research foundation upon which clinical nurses are expected to base their practice may be a latticework of rust.

The potential significance of basing nursing practice on unsound research is immense. In the U.S. each year, over one million (USDHHS, 1986) registered nurses (RNs) are responsible for 24-hour care, health education, and discharge planning for nearly 40 million hospital admissions (American Hospital Association, 1984). These RNs also provide education and nursing intervention for 273 million hospital outpatient visits (American Hospital Association, 1984); an additional 100,000 RNs provide similar services in non-hospital ambulatory care settings; and 20,000 nurse practitioners and midwives (USDHHS, 1986) provide primary care. Another 115,000 RNs (USDHHS, 1986) supervise the health of approximately 1.5 million residents of nursing homes and long-term care facilities (USDHHS, 1984), and 100,000 develop and evaluate health programs in schools, industries, and community health agencies (USDHHS, 1986).

Because nursing practice directly affects the health of millions of Americans, it is ethically imperative to identify and correct any deficiencies in the research upon which that practice is based. A major obstacle to such an evaluation has been the lack of an instrument for valid, reliable, and efficient data collection. The instrument to be tested in this study, the RAF (Appendix B), has been developed through an extensive review of evaluations of research in the health professions (RAF reference list, Appendix B). The RAF has been judged to have content validity through a series of reviews by a panel of nationally known experts in nursing research, biostatistics, and publishing. Preliminary work suggests that, after a training period, graduate nursing students may be able to complete the RAF with 90% reliability in two hours per article.

This project will extend reliability testing to a random sample of research articles published in five clinical nursing journals in 1986. This will enable us to refine the RAF and more firmly establish its reliability. Potential users of the RAF (or sections thereof) include reviewers of grant proposals, journal editors and manuscript reviewers, faculty and students in research courses, and researchers who need to screen published studies for meta-analytic research, as well as researchers who wish to conduct scientific studies of the methodological quality of research in any area of nursing.

This project will also provide a beginning evaluation of the scientific merit of research actually read by practicing nurses, and upon which they are expected to base their clinical practice. It will determine whether there are serious methodological flaws in this research, and, if so, will identify areas in need of correction. Nursing educators and researchers then can take specific actions to correct the errors, and practitioners and their clients can benefit from the findings of improved research. The data base created through this project also will provide a methodologically sound foundation for future statistical evaluation of changes in research quality over time.

C. Scientific Rationale

The assumptions underlying this study of the quality of published clinical nursing research are that 1) research is the appropriate basis for professional nursing practice (ANA, 1985; NLN, 1983, 1986); 2) research can vary in quality (see Significance, p.1, and RAF reference list, Appendix B); and 3) published articles that violate scientific principles and established standards for the conduct and reporting of research may jeopardize nursing practice (see Significance, p.2).

These principles and standards for the conduct and reporting of research (see RAF reference list, Appendix B) provide the scientific base for this study and are the foundation for construction of the instrument to be tested, refined, and used to evaluate published clinical nursing research in this project. These principles and standards, as they relate to published research articles (Selby, 1988, Appendix C; Tornquist, 1986;) are summarized below.

Overall, the reader of a research article should be able to follow the logical development of the research question, the methods used, the results, and the conclusions of the study. Though organizational style may vary, a good research article includes the background and rationale for the study; methodology used; results; discussion; implications; and limitations.

First, the article should introduce the reader to the problem area and the scope or significance of the problem. A synopsis of the literature should demonstrate a synthesis of ideas indicating the existing gaps in knowledge, shortcomings in previous research, and/or unmet needs which the author's research will remedy.

These introductory remarks should lead logically to the problem or purpose statement, which should be stated explicitly. From the purpose statement, it should be clear that the project is researchable and properly delimited. Any research questions or hypotheses should be derived directly from the purpose statement and fit within the stated scope of the study. The potential benefits or uses of information gained from answering the research questions should be stated succinctly, providing a coherent rationale to convince the reader that the study purpose is worthwhile.

High quality research is developed using conceptual, theoretical, or scientific frameworks or models within which the research questions are couched. If used, the concept, theory, model, or scientific base should be documented and identified; its relevance to the study should be made clear. Any assumptions should also be explicitly stated, appropriately based in theory, research, or universal truths, and be relevant and valid within the context of the study.

A clear description of the study's methodology is essential. Although journal length limitations may preclude a detailed description, at the very least an article should describe the research design, setting, population and sample, protection of human subjects, data collection instruments and procedures, and data analysis techniques.

The research design should be compatible with the problem statement, research questions, and/or hypotheses. The setting should be described in terms of the time period, location (overall geographic area and type of particular institution, unit, etc.), and other physical conditions relevant to the study. Limitations of the design and possible sources and directions of bias in the setting should be acknowledged.

The target population and sampling frame should be identified. The sampling technique used should be named and described, as should criteria for sample selection. Sample size should be noted and justified, and limitations of the sample, with possible sources and directions of bias, should be pointed out. If human subjects are involved, the article should provide evidence that informed consent was obtained.

Data collection instruments should be described with regard to how they measure the variables of interest; the level of measurement of the variables should be clear from the descriptions. The validity and reliability of each instrument should be addressed. The instruments should be congruent with the problem statement, research questions, hypotheses, and variables. Data collection procedures should be described in sufficient detail to enable others to replicate the study. Any problems with data collection which might affect the quality of the data should be discussed.

Data analysis methods should be adequate to address the research questions. Statistics, where used, should be identified and appropriate for the data. If non-standard statistical tests are used, or if statistics are used in an unusual manner, a rationale and references should be provided.

The presentation of the results should follow the description of the methodology. The final sample size, response rate, withdrawals and losses, and reasons for these should be provided if not already given in the methods section. Appropriate descriptive statistics for demographic characteristics of the sample should be presented in narrative or tabular form. If the sample is small, numbers should be presented along with percentages or proportions. The narrative for the major findings should provide a logical flow of information, highlighting important points and referring the reader to tables for detailed information where appropriate. Tables and figures must be understandable without reference to the text. Appropriate descriptive statistics for the major variables of interest should be provided, including differences between groups if applicable. The results of hypothesis testing, including the name and value of the statistical test and p-value, should be included. The results of secondary analyses and anecdotal data may be presented if space permits.

Only data presented in the results are appropriate for inclusion in the discussion section. In this section, the reader should find the answer to the questions, "So what? Why are the findings important?" It may be necessary to provide a brief summary of the major findings. The main focus should be on interpreting the meaning of the findings in view of the literature (research and/or theory), problems within the study, insights, observations, and opinions

(so noted as such). Appropriate generalizations about the data should be made. If not discussed in the methods section, limitations of the research design on the ability to draw conclusions, and limitations of the sample and setting on the ability to generalize should be noted. How or why the validity may be compromised should be discussed.

The authors should explain the implications of the study; these implications should relate to their findings, not information available before their research was undertaken. At the least, implications for future research should be discussed. Implications for practice or policy also should be included, even if the implications only caution practitioners and policy makers not to make changes because of the limitations of the study. If the study purports to test a theory or model, the implications for theory or model development also should be discussed.

The instrument to be tested and refined in this study is designed to measure adherence to the research principles and standards described above, and to identify specific weaknesses in research methodology and statistical use in published research articles. We will use this tool to assess research quality in a random sample of research reports in clinical nursing journals. Through our research methodology, we intend to demonstrate that it is possible to utilize and abide by the principles and standards of sound research when assessing adherence to these principles and standards in the published research of others.

D. Consultative Support

Consultation will be provided by a nationally known biostatistician with extensive experience in conducting research and teaching research methodology (see letter of agreement and biosketch, Appendix A).

E. Methods

1. Sample

The sampling frame consists of all 130 original research articles (those which report data collection and/or analysis) published in 1986 in five U.S. clinical nursing specialty journals (critical care, Heart & Lung; oncology, Oncology Nursing Forum; gerontology, Journal of Gerontological Nursing; maternal and child health, Journal of Obstetrical, Gynecological, and Neonatal Nursing;and community health, Public Health Nursing. The journals were identified from a listing (Swanson & McCloskey, 1986) of publishing practices of nursing journals (those whose audience was at least 50% nursing). The initial choice was the journal with the highest circulation in its specialty. Then, on the premise that journals with little research content are less likely to be read by clinicians who actively seek to base their practice on research, the proportion of research content was assessed. If the editor of the chosen journal did not report that at least 33% of its published content was research, the journal with the next highest circulation was selected until a journal meeting the research content criterion was identified. In one case, the journal with the highest circulation reported 33% research but, on examination, only 1 of 40 articles published in 1986 was research; therefore, the second journal, which met the research content criterion, was chosen. The combined circulation of the journals selected is over 130,000 (Swanson & McCloskey, 1986).

A random sample of 56 articles proportionately stratified by journal will be selected using a table of random numbers; 56 is the largest estimate required to ensure 95% confidence that the proportion of articles containing a methodological error is within .10 of the true proportion in the sampling frame. This sample size also provides a foundation for future statistically sound comparisons between this sample and others of equivalent size, to evaluate changes in research quality over time.

2. Instrument

The Research Assessment Form (RAF, Appendix B), synthesized from an extensive literature review (RAF reference list, Appendix B) will guide the evaluation of research articles in this retrospective nonexperimental study. The RAF provides a checklist for assessing the clarity and logic of the study's statement of purpose; use of the literature to justify the need for the study; clarity and relevance of the theoretical or conceptual framework; evidence of protection of human subjects; description, appropriateness, and acknowledgment of limitations of the research design and sampling method; evidence of sample size calculation; adequacy of statistical power; evidence of validity and reliability of data collection instruments; appropriateness of use and completeness of reporting of descriptive and inferential statistics; completeness of the title, abstract, and tables; and relevance of stated implications to the study. The RAF also asks for descriptive information such as the number of authors, their degrees, whether the research was grant funded, and the type of funding.

The RAF has been judged to have content validity through a series of reviews by a panel of nationally known experts (nurse researcher, research editor, and biostatistician). Preliminary reliability testing of 7 of the 8 sections of the RAF, using 5 selected articles, has shown 80 to 90% agreement between the reviews of graduate nursing students trained in use of the RAF and a "gold standard" decisive review by the principal investigator (PI) and biostatistical consultant. Both the PI and consultant have considerable experience in conducting research, teaching research methodology, and judging research quality.

The major purpose of this study is to extend reliability testing to all sections of the RAF, calculating intra-rater (test-retest) agreement as well as gold standard (inter-rater) agreement, and refining the RAF as needed. Experience in using the RAF to evaluate 56 articles in this study will enable us to develop a mathematical scoring system for the RAF. Such a system will make it possible for manuscript reviewers, students, and future researchers to make quantitative comparisons of research articles.

3. Procedures

All studies involving data abstraction are subject to human error. Study procedures designed to minimize the possibility of error in this project include abstractor training, reliability testing, and procedural quality controls. Also, the project is planned to provide a meaningful collegial research experience for the entire investigative team. The PI, coinvestigator (CI), research assistants (RAs), and biostatistics (BIOS) faculty consultant will collaborate in report writing, sharing authorship and thus responsibility for the integrity of the study.

Data will be collected by RAs (graduate students in nursing and biostatistics) supervised by the CI, with PI consultation. The CI will train RAs to use the RAF and will coordinate data collection. For reliability testing, the PI and BIOS faculty consultant will perform a "gold standard" decisive review of a subsample of articles. Working independently of one another, the RAs will each review these articles and re-review them in two weeks, without access to their previous review forms. Reliabilities (percent agreement) will be calculated intra-RA (test-retest) and between RAs and the gold standard (inter-rater). Further training, RAF revision, and/or repetitions of the above procedure will be instituted if reliabilities are below 80%. Thereafter, RAs, with access to consultation, will review separate subsamples of articles until all 56 are reviewed.

Several additional quality control checks will be implemented during data collection. The PI, with BIOS faculty consultation, will perform gold standard reviews of additional articles, the identities of which will not be disclosed to the RAs. The RAs will review these same articles and their RAFs will be examined for agreement with the gold standard. If reliabilities fall below 80%, further training and/or RAF revision will be undertaken. Throughout the study, the CI will examine all RA-completed RAFs for completeness. The PI and BIOS consultant will make definitive judgments concerning methodological errors and any areas questioned by the CI or RAs, and will provide feedback to the investigative team.

4. Data Analysis

Inter- and intra-rater reliability for the RAF will be calculated as described above. If the mean percentage of agreement is 80% or better, we will conclude that the RAF has demonstrated these types of reliability.

A scoring system will be developed for the RAF, based on examination of individual items and groups of items. We plan to calculate a summative "research quality score" for each of the major methodological areas considered by the RAF (e.g., sampling, research design, descriptive statistics, etc.). To be useful for inter-article comparisons, the score cannot include purely descriptive items (e.g., location of setting), cannot penalize articles for items which are not applicable (e.g., lack of randomization to groups, when the study is nonexperimental), and cannot multiply errors inequitably (e.g., when one item on the RAF reveals an error, and the following items clarify the type of error). At this point it is not clear whether the scores for the various areas can be summed for a total RAF score.

The research quality scores will be reported in terms of means and standard deviations, medians, modes, and percents (grouping the scores) for the overall sample. Comparisons of scores may be made according to categories of selected variables such as journal and specialty, number of authors, educational levels and professions of authors, type of funding, type of research design, etc.

Because of the intensive time and labor required for data collection, the sample size is calculated for descriptive statistics only, not for testing for significance of differences (e.g., between journals or other categories). Articles will be described and crosstabulated by journal and specialty, number of authors, educational levels and professions of authors, and type of funding. Evaluative data from the RAF will be grouped and described in terms of the

proportion of articles with errors related to clarity and logic of the study's statement of purpose; use of the literature to justify the need for the study; clarity and relevance of the theoretical or conceptual framework; evidence of protection of human subjects; description, appropriateness, and acknowledgment of limitations of the research design and sampling method; evidence of sample size calculation; adequacy of statistical power; evidence of validity and reliability of data collection instruments; appropriateness of use and completeness of reporting of descriptive and inferential statistics; completeness of the title, abstract, and tables; and relevance of stated implications to the study. Mean ($\pm$ s.d.) power for small, medium, and large effect sizes also will be calculated.

Results from this study can be generalized to research articles published in 1986 in the selected clinical nursing journals. This does not include all research relevant to clinical nursing, and our methodology relies on what is reported and published, not necessarily what is accomplished in research. Nevertheless, this study will provide a beginning evaluation of the research that is likely to be read by practicing clinicians, and will establish a methodologically sound foundation for future research.

5. Time Frame

Mo. 1-2: Obtain sample; conduct RA training sessions; revise RAF as needed

Mo. 3-5: Collect data; calculate reliabilities; revise RAF as needed

Mo. 6: Enter data on computer

Mo. 7-9: Analyze and interpret data; develop scoring system for RAF

Mo. 10-12: Prepare reports for publication and presentation

F. Human Subjects

Because public scrutiny is an expectation of publication, the authors whose works will be examined in this study can be considered to have given permission for their studies to be reviewed. Nevertheless, in order to prevent possible embarrassment to individuals, we will not report data at the individual author level. This proposal has Institutional Review Board approval (Appendix A).

G. Facilities and Resources

Facilities relevant to this project include the Health Sciences Library, which subscribes to the journals selected for this project; photocopying facilities; office space for the investigators; a conference room for RA training; one AT&T PC6300 (hard disk) with SPSS PC statistical software; access to VAX mainframe dataprocessing and printing facilities; and access to nationally known research consultants.

H. Collaborative Arrangements

Collaborative arrangements with other agencies are not required. The department chairs of participating investigators have approved this project.

I. References

Abdellah, F. G. (1970). Overview of nursing research 1955-1968, part 1. Nursing Research, 19, 6-17.

American Hospital Association (1984). Hospital statistics (1984 ed.). Chicago: Author.

American Nurses' Association (1985). Directions for nursing research: Toward the twenty-first century. Kansas City, MO: Author.

Beck, C. T. (1985). Theoretical frameworks cited in nursing research from January 1974-June 1985. Nurse Educator, 10(6), 36-38.

Benoliel, J. Q. (1983). Nursing research on death, dying, and terminal illness: Development, present state, and prospects. In H. H. Werley, J. J. Fitzpatrick, & R. L. Taunton (Eds.). Annual review of nursing research (pp. 101-130). New York: Springer Publishing Company.

Brown, J. S., Tanner, C. A., & Padrick, K. P. (1984). Nursing's search for scientific knowledge. Nursing Research, 33, 26-32.

Chalmers, T., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. Controlled Clinical Trials, 2, 31-49.

Cronenwett, L. R. (1982). Father participation in child care: A critical review. Research in Nursing and Health, 5, 63-72.

Diers, D., & Molde, S. (1979). Some conceptual and methodological issues in nurse practitioner research. Research in Nursing and Health, 2, 73-84.

Ellis, R. (1977). Fallibilities, fragments, and frames: Contemplation on 25 years of research in Medical-Surgical nursing. Nursing Research, 26, 177-182.

Felson, D. T., Cupples, L. A., & Meenan, R. F. (1984). Misuse of statistical methods in Arthritis and Rheumatism. Arthritis and Rheumatism, 27, 1018-1022.

Ganong, L. H. (1987). Integrative reviews of nursing research. Research in Nursing & Health, 10, 1-11.

Gortner, S. R. (1983). The history and philosophy of nursing science and research. Advances in Nursing Science, 5(2), 1-8.

Gortner, S. R., & Nahm, H. (1977). An overview of nursing research in the United States. Nursing Research, 26, 10-33.

Grant, M. M., & Padilla, G. V. (1985). An overview of cancer nursing research. Oncology Nursing Forum Supplement, 12(1), 28-39.

Henderson, V. (1957). An overview of nursing research. Nursing Research, 6(2), 61-71.

Highriter, M. E. (1977). The status of community health nursing research. Nursing Research, 26, 183-192.

Hill, A. B. (1971). Principles of medical statistics (9th ed.). New York: Oxford University Press.

Jacobsen, B. S., & Meininger J. C. (1985). The designs and methods of published nursing research: 1956-1983. Nursing Research, 34, 306-312.

Jacobsen, B. S., & Meininger J. C. (1986). Randomized experiments in nursing: The quality of reporting. Nursing Research, 35, 379-386.

Jones, S. L., & Jones, P. K. (1987). Detecting statistically significant differences. Journal of Psychosocial Nursing, 25, 38-42.

Lindsey, A. M. (1982). Phenomena and physiological variables of relevance to nursing, review of a decade of work: Part I. Western Journal of Nursing Research, 4 (4), 343-364.

Lindsey, A. M. (1983). Phenomena and physiological variables of relevance to nursing, review of a decade of work: Part II. Western Journal of Nursing Research, 5 (1), 41-63.

Loomis, M. E. (1985). Emerging content in nursing: An analysis of dissertation abstracts and titles: 1976-1982. Nursing Research, 34, 113-118.

McCloskey, J. C. (1981). The effects of nursing education on job effectiveness: An overview of the literature. Research in Nursing and Health, 4, 355-373.

Moustafa, N. G. (1985). Nursing research from 1977 to 1981. Western Journal of Nursing Research, 7, 349-356.

National League for Nursing (1983). Criteria for the appraisal of baccalaureate and higher degree degree programs (5th ed.). New York: Author.

National League for Nursing, Council of Baccalaureate and Higher Degree Programs (1986). Characteristics of graduate education in nursing (third revision). New York: Author.

O'Connell K. A. (1983). Nursing practice: A decade of research. In N. Chaska (Ed.). The nursing profession: A time to speak (pp. 183-201). New York: McGraw-Hill Book Company.

O'Connell, L., & Duffey, M. (1978). Research in nursing practice: Its present scope. In N. Chaska (Ed.). The nursing profession: Views through the mist (pp. 183-201). New York: McGraw-Hill Book Company.

O'Fallon, J. R., Dubey, S. D., Salsburg, D. S., Edmonson, J. H., Soffer, A., & Colton, T. (1978). Should there be statistical guidelines for medical research papers? Biometrics, 34, 687-695.

Pocock, S. J. (1983). Clinical trials: A practical approach. Chichester, England: John Wiley & Sons.

Pollock, S. E. (1987). Adaptation to chronic illness: Analysis of nursing research. Nursing Clinics of North America, 22 (3), 631-644.

Schor, S., & Karten, I. (1966). Statistical evaluation of medical journal manuscripts. The Journal of the American Medical Association, 195, 1123-1128.

Schwirian, P. M. (1984). Research on nursing students. In H. H. Werley, J. J. Fitzpatrick, & R. L. Taunton (Eds.). Annual review of nursing research (pp. 211-237). New York: Springer Publishing Company.

Selby, M. L. (1988). Research study guide. Chapel Hill, NC: Curriculum in Public Health Nursing, University of North Carolina School of Public Health.

Stehle, J. L. (1981). Critical care nursing stress: The findings revisited. Nursing Research, 30 (3), 182-186.

Swanson, E., & McCloskey, J. (1986). Publishing opportunities for nurses. A comparison of publishing policies and practices of 139 journals. Nursing Outlook, 34, 227-235

Tornquist, E. M. (1986). From proposal to publication: An informal guide to writing about nursing research. Menlo Park, CA: Addison-Wesley Publishing Co.

U. S. Department of Health & Human Services (1984). Trends in nursing and related care homes and hospitals. National Health Survey, Series 14, No. 30. Hyattsville, MD: U. S. Government Printing Office, No.(PHS)84-1825.

U. S. Department of Health & Human Services (1986). The registered nurse population. National sample survey of registered nurses, November, 1984. Springfield, VA: National Technical Information Center, Accession No. HRP-0906938.

Werley, H. H., Fitzpatrick, J. J., & Taunton, R. L. (Eds.). (1983-1986). Annual review of nursing research (Vols. 1-4). New York: Springer Publishing Company.

Appendix B

Research Assessment Form

(RAF)

SELBY RESEARCH ASSESSMENT FORM (RAF)
VERSION (May 19, 1988)

by

Maija L. Selby, Dr.P.H., R.N., C., C.P.N.P.
Curriculum in Public Health Nursing
School of Public Health
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27500-7400
Phone: (919) 966-1030

RESEARCH ASSESSMENT FORM (RAF) - VERSION May 19, 1988

ABSTRACTOR: Attach photocopied article. Use blue highlighter on article as directed

1. Abstractor Name ______________________ ___ ___ 1

2. Total minutes required for completion ______________ ___ ___/___ 3

3. Was consultation required? ___ 6

   1. yes
   2. no; SKIP TO SECTION I

4. With whom was consultation? ______________________ ___/___ 7

SECTION I (items 5 - 14): GENERAL INFORMATION

5. Article i.d. ____/____/____ 9

Code directly ____/____/____

6. Journal ____ 12

1. Heart & Lung
2. Issues in Mental Health Nursing
3. Journal of Gerontological Nursing
4. Journal of Obstetrical, Gynecological, and Neonatal Nursing
5. Oncology Nursing Forum
6. Public Health Nursing

7. Volume number ____/____ 13

Code directly ____/____

8. Issue number ____/____ 15

Code directly ____/____

9. Number of authors ____/____ 17

Code directly ____/____

10. Highest degree of first (or only) author ____ 19

1. less than baccalaureate
2. baccalaureate
3. master's
4. doctorate
9. information not provided

11. Profession of first (or only) author ____ 20

1. RN
2. MD
3. Pharmacist
4. Nutritionist
5. Psychologist
6. Statistician
7. Other; describe____________________
9. information not provided

12. Highest degree of any co-authors after first author ____ 21

1. less than baccalaureate
2. baccalaureate
3. master's
4. doctorate
8. N/A; no co-authors on this article
9. information not provided

13. Is any co-author (after first author) a nurse (RN)? ___ 22

1. yes
2. no
8. N/A; no co-authors on this article
9. information not provided

14. Grant funding acknowledged. **ABSTRACTOR: usually at bottom of first page with author listing, or at end in Acknowledgements section.** ___ 23

1. federal only
2. non-federal only
3. both federal and non-federal
4. none listed

SECTION II (items 15 - 25): PURPOSE, LITERATURE REVIEW, FRAMEWORK

15. Is purpose statement clear? ABSTRACTOR: highlight purpose statement in blue. ___ 24

1. explicitly stated in introductory paragraphs
2. not explicitly stated within introductory paragraphs, but implied therein
3. not in introductory paragraphs, but stated or implied later
4. totally unclear or absent

16. Stated purpose is to: ___ 25

1. describe existing situation
2. explain or test differences or relationships in existing situation
3. test an intervention
4. develop an instrument
5. other; describe ____________________
6. totally unclear or absent

17. Major focus of research ___ 26

1. clinical practice (patient/client oriented)
2. administration
3. education
4. other; describe ____________________

18. Is ANY literature cited in introduction or literature review? ___ 27

1. yes
2. no

19. Does cited literature justify need for this study? ___ 28

1. yes, clear justification
2. unclear or contradictory justification
3. no justification or no literature
4. other; explain ____________________

20. Total number of cited references in article (count reference list): ___/___/___ 29

Code directly: ___/___/___

21. Conceptual or theoretical framework ___ 32

1. separate section included
2. presented in introduction or literature review
3. not included or implied; SKIP TO 26

22. Identify conceptual or theoretical framework upon which study is based ___/___ 33

____

23. Is conceptual or theoretical framework described at all? ___ 35

1. yes
2. no

24. Is an attempt made to explain the relevance of framework to this study? ___ 36

1. yes
2. no

25. Is framework relevant to this study? ___ 37

1. yes, framework probably is relevant
2. unable to determine
3. framework clearly is not relevant or is contradictory to this study

SECTION III (items 26 - 44): SETTING, HUMAN SUBJECTS, SAMPLING

26. Is the setting for the study mentioned or implied? ___ 38

1. yes
2. no; SKIP TO 28

27. Specific setting stated (where they say their sample is from)

a. hospital inpatient setting ___ 39

1. yes
2. no

b. long term care facility ___ 40

1. yes
2. no

c. outpatient/ambulatory care facility (MD office, HMO, clinic, etc.) ___ 41

1. yes
2. no

d. public health agency ___ 42

1. yes
2. no

e. college or university ___ 43

1. yes
2. no

f. primary or secondary school ___ 44

1. yes
2. no

g. industrial or occupational setting ___ 45

1. yes
2. no

h. community or geographical area ___ 46

1. yes
2. no

i. home ___ 47

1. yes
2. no

j. other; describe ______________________________ ___ 48

1. yes
2. no

28. Was evidence of informed consent provided (including studies which look at medical records or other existing data about humans) ___ 49

1. yes, citing of institutional or review board approval
2. yes, other clear evidence of voluntary nature of the study was provided
3. unclear
4. no, evidence of voluntary nature was not provided
8. N/A (e.g., no involvement of human subjects)

29. Sampling method claimed to be used by authors ___ 50

1. probability (ABSTRACTOR: do not confuse random sampling with random assignment to groups)
2. nonprobability
3. not addressed
8. N/A; no sampling used SKIP TO 45

30. Sampling method actually described by authors ___ 51

1. probability (ABSTRACTOR: do not confuse random sampling with random assignment to groups)
2. nonprobability
3. not described sufficiently to categorize
4. not described at all

31. Is sampling frame mentioned (at all, anything)? ___ 52

1. yes
2. no

32. Total attempted sample size: ___ 53 ___/___/___/___/___

Code directly ___/___/___/___/___/___

666666. unclear or conflicting information; DESCRIBE____

____

999999. information not provided

33. Total completed sample size ___ 59 ___/___/___/___/___

Code directly ___/___/___/___/___/___

666666. unclear or conflicting information; DESCRIBE____

____

999999. information not provided

34. Total number of refusals, and/or withdrawals, and/or cases lost ___ 65 ___/___/___/___/___

Code directly ___/___/___/___/___/___

666666. unclear or conflicting information; DESCRIBE____

____

999999. information not provided

35. If intervention, is (are) the reason(s) given for withdrawals or cases lost? ___ 71

1. reported no (zero) withdrawals or cases lost; therefore no reasons needed
2. reported withdrawals and cases lost and reported reasons
3. reported withdrawals and cases lost but did not report reasons
4. did not report withdrawals and cases lost and did not report reasons
8. N/A; not an intervention

36. What rationale was provided for sample size selection? ___ 72

1. no rationale or "practical rationale" stated or implied
2. evidence of sample size calculation or reference to sample size tables
3. sample size not given

37. Were sample size calculations reported to be based on ___ 73

1. descriptive statistics?
2. other (inferential) statistical tests?
3. both descriptive and inferential statistics?
4. no mathematical calculations described

ABSTRACTOR: Answer items 38-41 only for studies whose main purpose is to describe a population (descriptive statistics). Otherwise, SKIP TO 42.

Was there evidence that the following information was used in calculating sample size?

38. Estimate of population proportion or variance of a major dependent variable? ___ 74

1. yes
2. unclear ______________________________
3. no

39. population size? ___ 75

1. yes
2. unclear ______________________________
3. no

40. desired level of confidence (e.g., 95% confident)? ___ 76

1. yes
2. unclear ______________________________
3. no

41. desired level of precision (e.g., "within .05 of the true proportion...")? ___ 77

1. yes
2. unclear ______________________________
3. no

ABSTRACTOR: Answer items 42-44 only for studies whose main purpose is to describe differences between groups (inferential statistics or statistical tests). Otherwise, SKIP TO 45.

Was there evidence that the following information was used in calculating sample size?

42. significance level (alpha)? ___ 78

1. yes
2. unclear ______________________
3. no

43. power? ___ 79

1. yes
2. unclear ______________________
3. no

44. effect size? ___ 80

1. yes
2. unclear ______________________
3. no

SECTION IV (items 45 - 51): RESEARCH DESIGN

45. Research design stated or implied by authors as being used ___ 81

1. experiment
2. quasiexperiment
3. nonexperiment
4. not stated or implied
5. other; describe________________________________

46. Was there manipulation of the independent variable (intervention)? ___ 82

1. yes
2. unclear
3. no

47. Did they use a control or comparison group? ___ 83

1. yes
2. unclear
3. no

48. Was there randomization to groups? ABSTRACTOR: Do not confuse with random sampling. ___ 84

1. yes
2. unclear
3. no

49. If authors claim that this is an experiment or that randomization was used, is the method of randomization described? ___ 85

1. yes, and the method described constitutes randomization
2. yes, but the method is not clear. EXPLAIN____________
   ____
3. yes, but the method described does NOT constitute randomization
   EXPLAIN____
4. no, the method of randomization is not described
8. N/A; randomization not claimed

50. Type of experiment or quasiexperiment ___ 86

1. 1 group measured at one point in time
2. 1 group before/after
3. 2 groups measured at one point in time
4. 2 groups before/after
5. other____
8. N/A; nonexperimental

51. Type of nonexperimental design ___ ___ 87

1. retrospective (past-oriented) but not case-control
2. retrospective case-control
3. cross-sectional or survey (present-oriented, at time of data collection)
4. prospective or cohort (future oriented, follows forward)
5. used retrospective data to follow prospectively
6. other ______________________________
7. not described clearly enough to categorize
8. not described at all
88. N/A; experiment or quasiexperiment

SECTION V (items 52 - 63): DATA COLLECTION, VALIDITY, RELIABILITY

52. Were data collection procedures described sufficiently to understand how major instruments were administered? (e.g., mail, in person, phone) ___ 89

1. yes
2. no; EXPLAIN ______________________________

____

53. Are operational definitions sufficient to understand how major variables are measured (stated or implied definitions)? ___ 90

1. All (or only) operational definitions are clear
2. Some operational definitions are clear, and some are not clear
3. All operational definitions are entirely unclear or confusing
   EXPLAIN ______________________________

____

4. Operational definitions are totally missing

54. Are instrument descriptions sufficient to understand how the instruments measure all major variables? ___ 91

1. All instrument descriptions are clear
2. Some instrument descriptions are clear, and some are not clear
3. All instrument descriptions are entirely unclear or confusing
   EXPLAIN ______________________________

____

4. Instrument descriptions are totally missing

55. Did author(s) state or imply that the instruments used have validity (any kind)? ___ 92

1. states or implies that each (or only) instrument used to collect original data has validity
2. states that $\geq$ one instrument has validity, but omits discussion or reference for validity of at least one instrument
3. completely omits statement regarding validity of instruments; SKIP TO 60
4. states or implies that validity of instruments was not assessed: SKIP TO 60
5. other; EXPLAIN ______________________________

____

ABSTRACTOR: Refer to validity chart to assess measures of validity claimed or implied for any instruments for items 56 - 59.

56. "face" validity ___ 93

1. did not claim
2. claimed and presented documentation consistent with face validity
3. claimed but presented documentation inconsistent with face validity;
   EXPLAIN ______________________________
4. claimed but presented only reference for documentation
5. claimed but did not present any documentation
6. implied but did not claim; EXPLAIN ______________

____

57. content validity ___ 94

1. did not claim
2. claimed and presented documentation consistent with content validity
3. claimed but presented documentation inconsistent with content validity; EXPLAIN ______________________
4. claimed but presented only reference for documentation
5. claimed but did not present any documentation
6. implied but did not claim; EXPLAIN ______________________

58. criterion-related validity (predictive) ___ 95

1. did not claim
2. claimed and presented documentation consistent with criterion-related validity
3. claimed but presented documentation inconsistent with criterion-related validity; EXPLAIN ______________________
4. claimed but presented only reference for documentation
5. claimed but did not present any documentation
6. implied but did not claim; EXPLAIN ______________________

59. construct validity ___ 96

1. did not claim
2. claimed and presented documentation consistent with construct validity
3. claimed but presented documentation inconsistent with construct validity; EXPLAIN ______________________
4. claimed but presented only reference for documentation
5. claimed but did not present any documentation
6. implied but did not claim; EXPLAIN ______________________

ABSTRACTOR: Refer to reliability chart to assess measures of reliability claimed or implied for any instruments for items 60 - 63.

60. Did author(s) state or imply that the instruments used have reliability (any kind)? ___ 97

1. states or implies that each (or only) instrument used to collect original data has reliability
2. states that $\geq$ one instrument has reliability, but omits discussion or reference for reliability of at least one instrument
3. completely omits statement regarding reliability of instruments; SKIP TO 64
4. states or implies that reliability of instruments was not assessed; SKIP TO 64
5. other; EXPLAIN ______________________

61. internal consistency (KR-20,21; coefficient or Cronbach alpha; split half; odd-even; Spearman-Brown prophecy) ___ 98

1. did not claim
2. claimed and presented documentation consistent with claim
3. claimed, presented appropriate documentation, but some reliability coefficients < .60
4. claimed but presented documentation inconsistent with claim; EXPLAIN ______________________________
5. claimed but presented only reference for documentation
6. claimed but did not present any documentation
7. implied but did not claim; EXPLAIN ______________________________

62. stability (test-retest, intra-rater) ___ 99

1. did not claim
2. claimed and presented documentation consistent with claim
3. claimed, presented appropriate documentation, but some reliability coefficients < .60
4. claimed but presented documentation inconsistent with claim; EXPLAIN ______________________________
5. claimed but presented only reference for documentation
6. claimed but did not present any documentation
7. implied but did not claim; EXPLAIN ______________________________

63. equivalence (inter-rater, parallel forms) ___ 100

1. did not claim
2. claimed and presented documentation consistent with claim
3. claimed, presented appropriate documentation, but some reliability coefficients < .60
4. claimed but presented documentation inconsistent with claim; EXPLAIN ______________________________
5. claimed but presented only reference for documentation
6. claimed but did not present any documentation
7. implied but did not claim; EXPLAIN ______________________________

SECTION VI (items 64 - 128): DATA ANALYSIS AND RESULTS

64. Is there a specific section in which methods of data collection and/or analysis are described? 101

1. yes
2. no

65. Where are methods of data analysis (type of statistics, NOT just the name of a computer package such as SPSS, SAS) reported? 102

1. in a specific section where methods of data analysis are described
2. elsewhere in the article
3. not reported anywhere in article, but computer package mentioned (which one? ________________)
4. completely omitted in article

ABSTRACTOR: Items 66-75 ask you to evaluate whether a reported descriptive statistic is used appropriately or inappropriately based on the level of measurement of the variable or rules for usage of descriptive statistics.

66. How is the mean (average) used? ABSTRACTOR: this measure is appropriate for interval/ratio data 103

1. always used with interval/ratio data
2. used with ordinal data at least once
3. used with nominal data at least once
4. unclear; level of measurement of variable unclear
8. N/A; mean not reported

67. How are the standard deviation, standard error, or variance used? ABSTRACTOR: these measures are appropriate for interval/ratio data 104

1. always used with interval/ratio data
2. used with ordinal data at least once
3. used with nominal data at least once
4. unclear; level of measurement of variable unclear
8. N/A; standard deviation, standard error, or variance not reported

68. Is a mean for a major independent, dependent, or co-variable reported without a measure of variability (standard deviation, standard error, or variance)? 105

1. yes
2. no
8. N/A; no mean reported

69. Is a mean for a sample characteristic reported without a measure of variability (standard deviation, standard error, or variance)? 106

1. yes
2. no
8. N/A; no mean reported

70. How is the median (mid-point) used? ABSTRACTOR: this measure is appropriate for numbered ordinal or interval/ratio data 107

1. always used with ordinal or interval/ratio data
2. used with nominal data at least once
3. unclear; level of measurement of variable unclear
8. N/A; median not reported

71. How is the range (minimum to maximum) used? ABSTRACTOR: this measure is appropriate for numbered ordinal or interval/ratio data 108

1. always used with ordinal or interval/ratio data
2. used with nominal data at least once
3. unclear; level of measurement of variable unclear
8. N/A; range not reported

72. Is the mode (most frequent score/response) used? ABSTRACTOR: this measure is appropriate for all data 109

1. mode is reported
2. mode not reported

73. Is percent or proportion used? ABSTRACTOR: these measures are appropriate for all data 110

1. percent or proportion is reported
2. percent or proportion not reported

74. Is the raw frequency (number) used? ABSTRACTOR: this measure is appropriate for all data 111

1. frequency is reported
2. frequency not reported

75. In any article that reports proportions or percents, does author commit the error of reporting percentages or proportions without referring to raw frequencies within a very small sample or subsample ($N \leq 20$)? 112

1. yes; Give EXAMPLE ____________________
2. no
8. N/A; no percent or proportion reported

76. Is sample described in terms of <u>all</u> major variables (independent, dependent) from problem statement, research question(s), and/or hypothesis(es)? ABSTRACTOR: article must report in text or table a measure of central tendency (unless nominal) and variability for each major variable 113

1. yes
2. unclear; unable to link results with research question(s) or hypothesis(es)
3. no; EXPLAIN ____________________

77. Are sample demographics described statistically? ABSTRACTOR: This requires reporting of means, medians, modes, ranges, frequencies, or percents for important characteristics of the sample; NOT just statements like "the average subject was age 13 and had been sick for 3 months" ___ 114

1. yes
2. no; EXPLAIN ____________________

78. If comparison of $\geq$ two groups is a major purpose, are descriptive statistics presented regarding any demographic or other possible confounding variables, separately for each group? ___ 115

1. yes
2. no; EXPLAIN ____________________
8. N/A; only one group

79. Is any descriptive statistic reported without referring to the variable it describes (e.g., "the mean was ____" in a context in which it is not clear whether it is the mean age, mean number of dogs, or mean anxiety score)? ___ 116

1. yes
2. no
8. N/A; no descriptive statistics reported

80. How many major hypotheses would be required (minimum) to analyze the major problem statement or purpose statement? ___/___ 117

Code directly ___/___

99. unclear; EXPLAIN ____________________

____

81. Were these major hypotheses tested for statistical significance (not necessarily appropriately)? ___ 119

1. yes; all were tested
2. partially; at least one was not tested
3. no; none were tested
4. unclear; EXPLAIN ____________________
8. N/A; no hypotheses stated or implied

82. How many statistical tests are reported or implied by a statement such as ($p \leq .05$), "were statistically significant," etc. ABSTRACTOR: not just the number of types of tests, but the total number of instances of applying a test; e.g., there might have been five Chi Square tests applied plus one t-test, totalling six different tests ___/___ 120

0. no statistical tests are reported
1. only one statistical test is reported
2. 2-5 tests are reported
3. 6-10 tests are reported
4. 11-20 tests are reported
5. more than 20 tests are reported
9. unable to determine; EXPLAIN ____________________

Title, author, and number of minutes to complete (for purposes of reintegrating with rest of questionnaire)
CODE NUMBER OF MINUTES

___ / ___ / ___ 122

ABSTRACTOR: Items 83-125 ask you to evaluate whether a reported statistical test is used appropriately or inappropriately, based on the assumptions underlying the use of each test. Use the Selby Chart, LeGault Instructions, Daniel or Remington and Schork textbooks, and/or your statistical consultant to make these evaluations. Do NOT count violations of assumption of normality or homoscedasticity as a violation. IF NO STATISTICAL TESTS ARE USED, SKIP TO 126.

USE THE FOLLOWING CODING SCHEME FOR QUESTIONS 83-102:

1. Clearly violates assumptions underlying use of test.
2. Unclear or unable to evaluate due to lack of information.
3. Clearly does <u>not</u> violate any assumptions for test used.
8. N/A; not used.

83. How is confidence interval used? ___ 125

84. How is Z-test (any kind) used? ___ 126

85. How is McNemar Test used? ___ 127

86. How is Fisher Exact Test used? ___ 128

87. How is Binomial Test used? ___ 129

88. How is Mann-Whitney U used? ___ 130

89. How is analysis of variance (ANOVA or F test) used? ___ 131

90. How is analysis of covariance (ANCOVA) used? ___ 132

91. How is Pearson Product Moment used? ___ 133

92. How is Spearman Rho used? ___ 134

93. How is Kendall Tau used? ___ 135

94. How is Median Test used? ___ 136

95. How is Sign Test used? ___ 137

96. How is Wilcoxon Test used? ___ 138

97. How is Cochran Q used? ___ 139

98. How is Multiple Regression used? ___ 140

99. How is Factor Analysis used? ___ 141

100. How is Kolmogorov-Smirnov Test used? ___ 142

101. How is Kruskal-Wallis Test used? ___ 143

102. How is other test used (except Chi Square or t-test)?
Name of test:____ 144

103. Is Chi Square used (any kind)? 145

1. yes
2. no; SKIP TO 108 IF CHI SQUARE IS NOT USED

104. Is Chi Square used when EXPECTED VALUE in any cell is 0, or >20% of EXPECTED VALUES in cells <5? 146

1. yes
2. no
3. insufficient information provided

105. Is Chi Square used with proportions instead of raw frequencies? 147

1. yes
2. no
3. insufficient information provided

106. Is Chi Square used with related (not independent) samples? 148

1. yes
2. no
3. insufficient information provided

107. Is Chi Square for Contingency Tables used when Chi Square for Goodness of Fit required? 149

1. yes
2. no
3. insufficient information provided

108. Is t-test used (any kind)? 150

1. yes
2. no; SKIP TO 113 IF T-TEST IS NOT USED

109. Is t-test for related samples used when independent sample t-test required? ABSTRACTOR: assume that independent sample t-test is used unless otherwise stated 151

1. yes
2. no
3. insufficient information provided
8. N/A; no independent samples

110. Is t-test for independent samples used when related sample t-test required? ABSTRACTOR: assume that independent sample t-test is used unless otherwise stated 152

1. yes
2. no
3. insufficient information provided
8. N/A; no related samples

111. Is t-test used for comparing more than 2 groups? (e.g., groups A,B, & C, and t-test is used 3 times for Groups A & B, A & C, B & C) 153

1. yes
2. no
3. insufficient information provided

112. Are multiple t-tests used (i.e., as in 111 above) without stating that alpha was reduced? 154

1. yes
2. no
8. N/A; no multiple tests used

113. Are repeated observations analyzed as independent (any test)? ABSTRACTOR: e.g., WHEN N=6 INFANTS BUT 300 CRYING EPISODES ARE ANALYZED OR N=10 PEOPLE MEASURED BEFORE AND AFTER AND N=20 SCORES ARE ANALYZED. 155

1. yes
2. no
3. insufficient information provided
8. N/A; no repeated observations

114. Does article report results of statistical testing without enabling reader to identify the statistical test used (either in methods or results)? ABSTRACTOR: e.g., "the results were not statistically significant" or "the hypothesis was rejected" or "the differences were significant" but you can not determine what test was used 156

1. yes
2. no
8. N/A; no statistical tests used

115. Is p-value (either exact, or "less than ____") missing (not reported) for at least one test? 157

1. yes
2. no
8. N/A; no statistical tests reported

116. Is there an "orphan p," i.e., a p-value is reported without reference to the name of the corresponding test statistic at least once? (Answer "NO" if the name of the corresponding test statistic clearly is given somewhere—e.g., statistic name in a table, and the p-value in the text. If it is impossible to link them, answer "YES.") 158

1. yes; at least one orphan p
2. no; no orphan p's
8. N/A; no statistical tests reported

117. Is alpha level (level of significance) reported or implied at least once? ___ 159

1. yes, explicitly reported at least once
2. implied by statement such as $p \leq .05$ at least once
3. not reported at all
8. N/A; *no statistical tests reported*

118. Based on the statistic alpha and p-value, does the author wrongly "reject" or "fail to reject" the null hypothesis? ___ 160

1. yes, "rejects" alternate hypotheses, at least once
2. yes, wrongly rejects based on values provided, at least once
3. yes, does both 1 & 2, at least once each
4. yes, other (describe) ____________________
5. no, clearly makes correct decision for all tests
6. unable to evaluate because of lack of information on statistic or p-value for at least one test

119. Does author state that failure to reject any of the null hypotheses proves that the null is true? ___ 161

1. yes
2. no

120. Are sufficient data presented to allow you to determine POWER for the tests for the major purpose, research question(s), or hypothesis(es)? ABSTRACTOR: Refer to instructions and Cohen tables ___ 162

1. yes; *SKIP TO 122*
2. no

121. If sufficient data are not presented, is one of the following items missing?

USE THE FOLLOWING CODING SCHEME FOR QUESTIONS 121a-d:

1. yes
2. no

a. Sample size ___ 163

b. Name of the statistical test ___ 164

c. Alpha value ___ 165

d. Effect size ___ 166

122. Power for statistical test for first major hypothesis (from Cohen tables) CODE ONLY REAL POWER

Name of test:______________________________

Value of test:______________________________

Sample size:______________________________

Alpha value:______________________________

Effect size:______________________________

Real power:______________________________ ___/___/___ 167

IF EFFECT SIZE IS NOT MISSING, SKIP TO 123

If effect size for first major hypothesis is missing:

a. Power for large effect

Code directly ___/___/___ ___/___/___ 170

999. unable to compute
777. hypothesis stated but not tested (e.g., has 3 hypotheses stated but tests only first & second)

b. Power for medium effect

Code directly ___/___/___ ___/___/___ 173

999. unable to compute
777. hypothesis stated but not tested (e.g., has 3 hypotheses stated but tests only first & second)

c. Power for small effect

Code directly ___/___/___ ___/___/___ 176

999. unable to compute
777. hypothesis stated but not tested (e.g., has 3 hypotheses stated but tests only first & second)

123. Power for statistical test for second major hypothesis (from Cohen tables) CODE ONLY REAL POWER

Name of test:______________________________

Value of test:______________________________

Sample size:______________________________

Alpha value:______________________________

Effect size:______________________________

Real power:______________________________ ___/___/___ 179

IF EFFECT SIZE IS NOT MISSING, SKIP TO 124

If effect size for <u>second major</u> hypothesis is missing:

a. Power for large effect

Code directly ____/____/____ ____ ____/____ 182

777. hypothesis stated but not tested
888. N/A; no second hypothesis
999. unable to compute

b. Power for medium effect

Code directly ____/____/____ ____/____/____ 185

777. hypothesis stated but not tested
888. N/A; no second hypothesis
999. unable to compute

c. Power for small effect

Code directly ____/____/____ ____/____/____ 188

777. hypothesis stated but not tested
888. N/A; no second hypothesis
999. unable to compute

124. Does this study FAIL TO REJECT a <u>major</u> null hypothesis? ____ 191

1. yes
2. unclear; EXPLAIN____________________________________
3. no

125. If study fails to reject a <u>major</u> null hypothesis, what is the POWER for this hypothesis? ABSTRACTOR: if more than one, choose first one reported that you can calculate

CODE ONLY REAL POWER

Name of test:____________________________________

Value of test:____________________________________

Sample size:____________________________________

Alpha value:____________________________________

Effect size:____________________________________

Real power:____________________________________ ____/____/____ 192

888. N/A; does not fail to reject null hypothesis

IF EFFECT SIZE IS NOT MISSING, SKIP TO 126

If effect size is missing for major null hypothesis that was not rejected:

a. Power for large effect

Code directly ____/____/____ ____/____/____ 195

999. unable to compute

b. Power for medium effect

Code directly ____/____/____ ____/____/____ 198

999. unable to compute

c. Power for small effect

Code directly ____/____/____ ____/____/____ 201

999. unable to compute

SECTION VII (Items 126 - 136): IMPLICATIONS

126. If no statistical significance is found with a small sample or subgroup, does author place considerable confidence in the lack of significance? ABSTRACTOR: e.g., does the author make a "big deal" about there being no difference or no relationship, and not refer to the fact that the small sample may have caused it? ___ 204

1. yes
2. no

127. Were results generalized beyond the sampling frame? ___ 205

1. yes, clearly beyond sampling frame; EXPLAIN ____________________
2. unclear; EXPLAIN ____________________
3. no
8. N/A; no sample

128. Were results generalized beyond sample? ___ 206

1. yes; EXPLAIN ____________________
2. unclear; EXPLAIN ____________________
3. no
8. N/A; no sample

129. Were ANY limitations of sampling stated? ___ 207

1. yes
2. no

130. List serious limitations overlooked, if any (confounding factors) ___/___ 208

____
____
(to be coded later)

131. Are any conclusions INCONSISTENT with results? ___ 210

1. yes; EXPLAIN ____________________
2. no

132. Are there any stated implications (research, theory, practice, etc.)? ___ 211

1. yes
2. no; SKIP TO 137

133. Are implications stated for future research? ___ 212

1. yes; EXPLAIN ____________________
____
2. unclear ____________________
3. no

134. Are implications stated for theory? ___ 213

1. yes; EXPLAIN ______________________________
____
2. unclear ______________________________
3. no

135. Are implications stated for practice or policy? ___ 214

1. yes; EXPLAIN ______________________________
____
2. unclear ______________________________
3. no

136. Do any of this study's findings CONTRADICT, or are they inconsistent with, stated implications? ___ 215

1. yes; EXPLAIN ______________________________
____
2. no

SECTION VIII (Items 137 - 144): TABLES, FIGURES, ABSTRACT, TITLE

137. Is title of article consistent with report? ___ 216

1. yes
2. no; EXPLAIN ____________________

____

138. Does article include any tables or figures? ___ 217

1. yes
2. no; SKIP TO 141

139. Do the table/figure titles sufficiently identify the contents of the tables/figures (could they stand alone in a table of contents)? ___ 218

1. yes, all table/figure titles correctly describe contents of table/figure
2. no, at least 1 table/figure title does not correctly describe contents of table/figure;
   EXAMPLE ____________________

140. Is content within tables/figures understandable without reference to the text? ___ 219

1. yes, all tables/figures are clearly understandable without reference to text
2. no, at least 1 table/figure lacks appropriate labeling;
   EXAMPLE ____________________

____

141. Does article include an abstract? ___ 220

1. yes
2. no; OMIT 142-144

142. Does abstract include purpose of study? ___ 221

1. yes, includes purpose
2. no, does not include purpose
3. other comment ____________________

143. Does abstract include summary of methods? ___ 222

1. yes
2. no

144. Does abstract include main results? ___ 223

1. yes
2. no

ABSTRACTOR: YOU HAVE REACHED THE END. THANK YOU FOR YOUR EFFORT!

Table 1

Guide to Techniques for Determining Instrument Validity

Type	Key Question	Technique	Comments
• Face	Does the instrument appear to be logical?	Researcher decision	• Alone, this is not a legitimate form of validity
Content	Are the individual items representative of the concept?	Panel of experts	A. Weakest form of true validity. B. Relatively simple to achieve.
Criterion-Related	Is there a high correlation between this measure and another measure of the same concept?	Correlation coefficient (between the 2 measures)	A. Difficult to decide on appropriate criterion. B. If another valid measure is already available, is there a legitimate reason for not using the other measure?
Construct	Does the instrument truly measure the concept and not something else? Note: Construct validity is the highest form of validity	1. Known-groups	A. Assumption is made about characteristics of the "known groups." B. Least difficult of 3 construct validity techniques
		2. multitrait-multi-method matrix	A. Requires 2 or more methods and concepts. B. Requires some statistical sophistication.
		3. Factor analysis	A. Requires some statistical sophistication.

Table 2

Guide to Techniques for Determining Instrument Reliability

Type	Key Questions	Technique	Comments
Equivalence	1. Do different researchers using the same instrument measure the same characteristic?	Inter-rater reliability	Used when more than one researcher using an instrument that requires subjective judgement.
	2. Do different instruments *measure the same* characteristic?	Parallel forms reliability	Used when it is advisable to have different items for the pretest and post-test.
Stability	Will instrument give the same results on repeat administration?		
	1. Instrument to be completed by researcher.	Intra-rater reliability	Not appropriate if data being measured are not stable over time.
	2. Instrument to be completed by *subjects*.	Test-retest reliability	Short period of time between first and second administration may give falsely high reliability because person completing instrument may simply remember earlier responses.
Internal Consistency	Do the sub-parts measure the *same characteristic*?	1. For Likert-type scale: A. Coefficient alpha B. Split half or odd/even with correction by Spearman-Brown Prophecy 2. For right/wrong questions: A. Kuder-Richardson 20 (KR20) B. Split half or odd/even with co ection by . . earman-Brown Prophecy	This type of reliability is not appropriate unless the various items are intended to measure the same concept.
	Are individuals respondin; consistently?	Agreement of responses between 2 items that ask for the same information.	Used when researcher is concerned about possibility that a particular item will not be measured accurately.

Table

Guide to Selected Inferential Statistics for Testing for Differences between Groups

Type(s) of group(s)	Level of Measurement of Dependent Variable: Nominal	Ordinal	Interval/Ratio
1 sample (comparing distribution of variable to a hypothesized distribution)	1. Chi Square for Goodness of fit	1. Sign Test 2. Wilcoxon Matched Pairs Signed-Rank Test	Means: 1. unknown population variance: one-sample t-test, df = n-1 2. known population variance: one-sample Z-test. Proportions: 1. one-sample Z-test for proportions
2 independent groups	1. Chi Square Tests 2. Fisher Exact Test	1. Median Test 2. Mann-Whitney U (Rank Sum) Test	Means: 1. two-sample t-test for independent samples. $df = n_1 + n_2 - 2$ Proportions: 1. two-sample Z-test for proportions
>2 independent groups	1. Chi Square Tests	1. Extension of Median Test 2. Kruskal-Wal... Test	Means: 1. one-way analysis of variance (one-way ANOVA)
1 group before/after, or 2 related groups (matched pairs)	1. McNemar Test	1. Wilcoxon Matched Pairs Signed-rank Test	Difference scores: 1. paired t-tests for paired data (t-test for related samples) df = n-1 where n = number of pairs
1 group with > 2 repeated measures, OR > 2 related groups (matched triplets, etc.)	1. Cochran Q	1. Friedman Test	Repeated measure analysis

- continued -

Guide to Selected Inferential Statistics - continued

	Level of Measurement of Dependent Variable		
Type(s) of group(s)	Nominal	Ordinal	Interval/Ratio
2 groups before/after	1. Adaptation of data for Chi Square Tests 2. Adaptation of data for Fisher Exact Test	1. Adaptation of data for Mann-Whitney U (if scores can be subtracted)	<u>Mean changes</u>: 1. two-sample t-test for independent samples. $df = n_1 + n_2 - 2$ 2. one-way analysis of variance <u>Means</u>: 1. analysis of covariance
>2 groups before/after	1. Adaptation of data for Chi Square Tests	1. Adaptation of data for Kruskal-Wallis test (if scores can be subtracted)	<u>Mean changes</u>: 1. one-way analysis of variance <u>Means</u>: 1. analysis of covariance

NOTE: *Lower level tests may be used for higher level data, but some loss of information may result.*

This chart was developed with the assistance of Ronald Forthofer, Ph.D., Professor of Biometry, University of Texas School of Public Health.

Appendix C

Research Assessment Form References

RAF References

Abdellah, F. G., & Levine E. (1979). Better patient care through nursing research (2nd ed.). New York: Macmillan Publishing Co., Inc.

Altman, D. G. (1980). Statistics and ethics in medical research. Collecting and screening data. British Medical Journal, 281, 1399-1401.

Altman, D. G. (1980). Statistics and ethics in medical research. Misuse of statistics is unethical. British Medical Journal, 282, 1182-1184.

Altman, D. G. (1980). Statistics and ethics in medical research. Study design. British Medical Journal, 281, 1267-1269.

Altman, D. G. (1980). Statistics and ethics in medical research. V. Analysing data. British Medical Journal, 281, 1473-1475.

Altman, D. G. (1980). Statistics and ethics in medical research. VI. Presentation of results. British Medical Journal, 281, 1542-1544.

Altman, D. G. (1980). Statistics and ethics in medical research. VII. Interpreting results. British Medical Journal, 281, 1612-1614.

Altman, D. G. (1980). Statistics and ethics in medical research. VIII. Improving the quality of statistics in medical journals. British Medical Journal, 282, 44-47.

Bailar III, J. C., Louis, T. A., Lavori, P. W., & Polansky, M. (1984). A classification for biomedical research reports. The New England Journal of Medicine, 311, 1482-1487.

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 9, 391-401.

Brown. C.G., Kelen, G. D., Moser, M., Moeschberger, M. L., & Rund, D. A. (1985). Methodology reporting in three acute care journals: Replication and reliability. Annals of Emergency Medicine, 14, 986-991.

Brown, J. S., Tanner, C. A., & Padrick, K. P. (1984). Nursing's search for scientific knowledge. Nursing Research, 33, 26-32.

Bunce, H. III, Hokanson, J. A., & Weiss, G. B. (1980). Avoiding ambiguity when reporting variability in biomedical data. American Journal of Medicine, 69, 8-9.

Burns, N., & Grove S. K. (1987). The practice of nursing research conduct, critique and utilization. Philadelphia: W. B. Saunders Company.

Chalmers, T., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. Controlled Clinical Trials, 2, 31-49.

Chase, L. J., & Tucker, R. K. (1975). A power-analytic examination of contemporary communication research. Speech Monographs, 42(3), 29-41.

Clarfield, A. M., & Friedman, R. (1985). Survey of the age structure of "age-relevant" articles in four general medical journals. Journal of the American Geriatrics Society, 33, 773-778.

Cobb, A. K., & Hagemaster, J. N. (1987). Ten criteria for evaluating qualitative research proposals. Journal of Nursing Education, 26 (4), 138-143.

Cohen, J. (1962). The statistical power of abnormal-social psychological research. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Colditz, G. A., & Emerson, J. D. (1985). The statistical content of published medical research: some implications for biomedical education. Medical Education, 19, 248-255.

Colton, T. (1974). Statistics in medicine. Boston: Little, Brown and Company.

Cook, T. D., & Leviton, L. C. (1983). Reviewing the literature: A comparison of traditional methods with meta-analysis. In R. J. Light (Ed.). Evaluation studies review annual (pp. 59-82). Beverly Hills: Sage Publications.

Cook, T. D.,. & Campbell, D. T. (1979). Quasi-experimentation. Design & analysis issues for field settings. Boston: Houghton Mifflin Company.

Cooper, H. M., & Arkin, R. M. (1983). On quantitative reviewing. In R. J. Light (Ed.). Evaluation studies review annual (pp. 167-172). Beverly Hills: Sage Publications.

Cooper, H. M., & Rosenthal, R. (1983). Statistical versus traditional procedures for summarizing research findings. In R. J. Light (Ed.). Evaluation studies review annual (pp. 83-90). Beverly Hills: Sage Publications.

Daniel, W. W. (1978). Applied nonparametric statistics. Boston: Houghton Mifflin Company.

DerSimonian, R., Charette, L. J., McPeek, B., & Mosteller, F. (1982). Reporting on methods in clinical trials. The New England Journal of Medicine, 306, 1332-1337.

Devine, E. C., & Cook, T. D. (1983). A meta-analytic analysis of effects of psychoeducational interventions on length of postsurgical hospital stay. Nursing Research, 32, 267-274.

Edlund, M. J., Craig, T. J., & Richardson, M. A. (1985). Informed consent as a form of volunteer bias. American Journal of Psychiatry, 142, 624-627.

Eifert, G. H., & Klant M. (1983). Inadequate presentation of behavioral measures of fear in the major journals. Journal of Behavior Therapy and Experimental Psychiatry, 14, 219-221.

Elenbaas, J. K., Cuddy, P. G., & Elenbaas, R. M. (1983). Evaluating the medical literature, part III: Results and discussion. Annals of Emergency Medicine, 12, 679-686.

Emerson, J. D., & Colditz, G. A. (1983). Use of statistical analysis in The New England Journal of Medicine. The New England Journal of Medicine, 309, 709-713.

Emerson, J. D., McPeek, B., & Mosteller, F. (1984). Reporting clinical trials in general surgical journals. Surgery, 95, 572-579.

Feinstein, A. R. (1978). Clinical biostatistics. XLIV. A survey of the research architecture used for publications in general medical journals. Clinical Pharmacology and Therapeutics, 24(1), 117-125.

Feinstein, A. R., & Horwitz, R. I. (1982). Double standards, scientific methods, and epidemiologic research. The New England Journal of Medicine, 307, 1611-1617.

Felson, D. T., Cupples, L. A., & Meenan, R. F. (1984). Misuse of statistical methods in Arthritis and Rheumatism. Arthritis and Rheumatism, 27, 1018-1022.

Fitz-Gibbon, C. T., & Morris, L. L. (1987). How to analyze data. Beverly Hills: Sage Publications.

Fletcher, R. H., & Fletcher, S. W. (1979). Clinical research in general medical journals. A 30-year perspective. The New England Journal of Medicine, 301, 180-183.

Freiman, J. A., Chalmers, T. C., Smith, Jr, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. The New England Journal of Medicine, 299, 690-694.

Ganong, L. H. (1987). Integrative reviews of nursing research. Research in Nursing & Health, 10, 1-11.

Gentry, M., & Shulman, A. D. (1985). Survey of sampling techniques in widowhood research, 1973-1983. Journal of Gerontology, 40, 641-643.

Glantz, S. A. (1980). Biostatistics: How to detect, correct and prevent errors in the medical literature. Circulation, 61, 1-7.

Glantz, S. A. (1987). Primer of biostatistics (2nd ed.). New York: McGraw-Hill Book Company.

Glass, G. V. (1980). Summarizing effect sizes. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 13-32). San Francisco: Jossey-Bass Inc., Publishers.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage Publications.

Gore, S. M., Jones, I. G., & Rytter, E. C. (1977). Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. British Medical Journal, 1, 85-87.

Grundner, T. M. (1986). Informed consent: A tutorial. Owings Mills, MD: National Health Publishing.

Haines, S. J., (1981). Six statistical suggestions for surgeons. Neurosurgery, 9, 414-418.

Harlem, O., Huth, E. J., Lock, S. P., Munro, I., Relman, A., Riis, P., Robinson, R., Sherrington, A., Southgate, M. T., & Vartiovaara, I. (1982). Uniform requirements for manuscripts submitted to biomedical journals. Annals of Internal Medicine, 96, 766-771.

Hartshorn, J. C. (1987). Research-based practice: The need for, use, and reporting of instrument reliability and validity. Heart & Lung, 16, 100-101.

Hayman, L. L. (1987). Fatal flaws. Nursing Research, 36 (5), 267.

Haynes, D. H. (1983). Contribution of statistics to ethics of science. American Journal of Physiology, 244, 3-5.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando: Academic Press, Inc.

Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. (1987). How to measure attitudes (2nd ed.). Beverly Hills: Sage Publications.

Hill, A. B. (1971). Principles of medical statistics (9th ed.). New York: Oxford University Press.

Hopkins, K. D. (1973). Preventing the number one misinterpretation of behavioral research, or how to increase statistical power. The Journal of Special Education, 7(1), 103-107.

Horwitz, R. I., & Feinstein, A. R. (1979). Methodologic standards and contradictory results in case-control research. The American Journal of Medicine, 66, 556-564.

Hovell, M. F. (1982). The experimental evidence for weight-loss treatment of essential hypertension: A critical review. American Journal of Public Health, 72, 359-368.

Huck, S. W., Cormier, W. H., & Bounds, Jr., W. G. (1974). Reading statistics and research. New York: Harper & Row, Publishers.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis. Cumulating research findings across studies. Beverly Hills: Sage Publications.

Jackson, G. B. (1983). Methods for integrative reviews. In R. J. Light (Ed.). Evaluation studies review annual (pp. 133-156). Beverly Hills: Sage Publications.

Jacobsen, B. S., & Meininger, J. C. (1985). The designs and methods of published nursing research: 1956-1983. Nursing Research, 34, 306-312.

Jacobsen, B. S., & Meininger, J. C. (1986). Randomized experiments in nursing: The quality of reporting. Nursing Research, 35, 379-386.

Jones, S. L., & Jones, P. K. (1987). Detecting statistically significant differences. Journal of Psychosocial Nursing, 25, 38-42.

Joyce, K. (1985). Critical evaluation of nephrology nursing research. ANNA Journal, 12(1), 39-40.

Kerlinger, F. N. (1973). Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart and Winston, Inc.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials, 2, 93-113.

LaPorte, R. E., & Cresanta, J. L. (1980). Research trends in the American Journal of Epidemiology. American Journal of Epidemiology, 111, 137-141.

Larson, D. B., Pattison, E. M., Blazer, D. G., Omram, A. R., & Kaplan, B. H. (1986). Systematic analysis of research on religious variables in four major psychiatric journals, 1978-1982. American Journal of Psychiatry, 143 (3), 329-334.

Lee, K. L., McNeer, F., Starmer, C. F., Harris, P. J., & Rosati, R. A. (1980). Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61, 508-515.

Leviton, L. C., & Cook, T. D. (1983). What differentiates meta-analysis from other forms of review? In R. J. Light (Ed.). Evaluation studies review annual (pp. 173-178). Beverly Hills>: Sage Publications.

Levy, P. S., & Lemeshow, S. (1980). Sampling for health professionals. Belmont, CA: Lifetime Learning Publications.

Light, R. J. (Ed.). (1983). Evaluation studies review annual. Beverly Hills: Sage Publications.

Light, R. J., & Pillemer, D. B. (1983). Numbers and narrative: Combining their strengths in research reviews. In R. J. Light (Ed.). Evaluation studies review annual (pp. 33-58). Beverly Hills: Sage Publications.

Light, R. J., & Pillemer, D. B. (1984). Organizing a reviewing strategy. Summing up. The science of reviewing research (pp. 13-103). Cambridge, MA: Harvard University Press.

Louis, T. A., Fineberg, H. V., & Mosteller, F. (1985). Findings for public health from meta-analysis. In L. Breslow, J. E. Fielding, & L. B. Lave (Eds.). Annual Review of Public Health (pp. 1-20). Palo Alto: Annual Reviews Inc.

Lynn, M. R. (1985). Reliability estimates: Use and disuse. Nursing Research, 34 (4), 254-256.

Mainland, D. (1963). The significance of "nonsignificance". Clinical Pharmacology and Therapeutics, 4, 580-586.

Mosteller, F. (1979). Problems of omission in communications. Clinical Pharmocology and Therapeutics, 25, 761-764.

Mosteller, F., Gilbert, J. P., & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: Agenda for the editor. Controlled Clinical Trials, 1, 37-58.

Munro, B. H., Visintainer, M. A., & Page, E. B. (1986). Statistical methods for health care research. Philadelphia: J. B. Lippincott Company.

O'Fallon, J. R., Dubey, S. D., Salsburg, D. S., Edmonson, J. H., Soffer, A., & Colton, R. (1978). Should there be statistical guidelines for medical research papers? Biometrics, 34, 687-695.

O'Flynn, A. I. (1982). Meta-analysis. Nursing Research, 31 (5), 314-316.

Pagano, R. R. (1981). Understanding statistics in the behavioral sciences. St. Paul: West Publishing Company.

Pillemer, D. B., & Light, R. J. (1980). Benefitting from variation in study outcomes. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 1-12). San Francisco: Jossey-Bass Inc., Publishers.

Pocock, S. J. (1983). Clinical trials: A practical approach. Chichester, England: John Wiley & Sons.

Polit, D. R., & Hungler, B. P. (1987). Nursing research principles and methods (3rd ed.). Philadelphia: J. B. Lippincott Company.

Prescott, P. A. (1987). Multiple regression analysis with small samples: Cautions and suggestions. Nursing Research, 36 (2), 130-133.

Rabins, P. V., Rovner, B. W., Larson, D. B., Burns, B. J., Prescott, C., & Beardsley, R. S. (1987). The use of mental health measures in nursing home research. Journal of the American Geriatric Society, 35, 431-434.

Reed III, J. F., & Slaichert, W. (1981). Statistical proof in inconclusive 'negative' trials. Archives of Internal Medicine, 141, 1307-1310.

Remington, R. D., & Schork, M. A. (1970). Statistics with applications to the biological and health sciences. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rosenthal, R. (1980). Summarizing significance levels. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 33-46). San Francisco: Jossey-Bass Inc., Publishers.

Rosenthal, R., & Rubin, D. B. (1980). Summarizing 345 studies of interpersonal expectancy effects. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 79-96).

Rosenthal, R., & Rubin, D. B. (1983). Comparing effect sizes of independent studies. In R. J. Light (Ed.). Evaluation studies review annual (pp. 235-239). Beverly Hills: Sage Publications.

Schor, S., & Karten, I. (1966). Statistical evaluation of medical journal manuscripts. The Journal of the American Medical Association, 195, 1123-1128.

Selby, M. L. (1988). Research study guide. Chapel Hill, NC: Curriculum in Public Health Nursing, University of North Carolina School of Public Health.

Sheehan, T. J. (1980). The medical literature. Let the reader beware. Archives of Internal Medicine, 140, 472-474.

Smith, M. C., & Naftel, D. C. (1984). Meta-analysis: A perspective for research synthesis. Image, 16, 9-13.

Smith, M. L. (1980). Integrating studies of psychotherapy outcomes. In R. Rosenthal (Ed.). Quantitative assessment of research domains. New directions for methodology of social and behavioral science (pp. 47-62). San Francisco: Jossey-Bass Inc., Publishers.

Soeken, K. L. (1985). Critiquing research. Steps for complete evaluation of an article. AORN Journal, 41, 882-893.

Stehle, J. L. (1981). Critical care nursing stress: The findings revisited. Nursing Research, 30, 182-186.

Stempel, L. E. (1982). Eenie, meenie, minie, moe ... What do the data really show? American Journal of Obstetrics and Gynecology, 144 (7), 745-752.

Strauss, S. (1969). Guidelines for analysis of research reports. The Journal of Educational Research, 63, 165-169.

Tanner, C. A. (1987). Evaluating research for use in practice: Guidelines for the clinician. Heart & Lung, 16 (4), 424-431.

Tornquist, E. M. (1986). From proposal to publication: An informal guide to writing about nursing research. Menlo Park, CA: Addison-Wesley Publishing Co.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

Volicer, B. J. (1984). Multivariate statistics for nursing research. Orlando: Grune & Stratton, Inc.

Wallenstein, S., Zucker, C. L., & Fleiss, J. L. (1980). Some statistical methods useful in circulation research. Circulation Research, 47(1), 1-9.

Weech, A. A. (1974). Statistics: Use and misuse. Australian Paediatric Journal, 10, 328-333.

Whiting-O'Keefe, Q. E., Henke, C., & Simborg, D. W. (1984). Choosing the correct unit of analysis in medical care experiments. Medical Care, 22, 1101-1114.

Williams, B. (1978). A sampler on sampling. New York: John Wiley & Sons.

Williams, M. E., & Retchin, S. M. (1984). Clinical geriatric research: Still in adolescence. Journal of the American Geriatrics Society, 32 (11), 851-857.

Wilson, H. S. (1985). Research in nursing. Menlo Park, CA: Addison-Wesley Publishing Company.

Wortman, P. M. (1983). Meta-analysis: A validity perspective. In R. J. Light (Ed.). Evaluation studies review annual (pp. 157-166). Beverly Hills: Sage Publications.

Young, M. J., Bresnitz, E. A., & Strom, B. L. (1983). Sample size nomograms for interpreting negative clinical studies. Annals of Internal Medicine, 99, 248-251.

Appendix D

Calculations for Sample Size, Confidence, and Precision

Appendix D

Calculations for Sample Size, Confidence, and Precision

Part I. Calculations for Sample Size[a]

These calculations were based on the researchers' desire to determine the proportion of articles (out of N = 130) that had one or more major methodological errors, given

$$n = \frac{t^2 p q}{d^2}$$

where N = 130 (sampling frame size)

$t = 1.96$ (i. e., 95% confidence)

$p = .5$ ("worst case" proportion to have ≥ 1 major methodological error)

$q = 1 - p = .5$ (for calculating proportion variance)

$d = .1$ (desired precision)

$$n = \frac{(1.96)^2 (.5)(.5)}{(.1)^2}$$

$$= 96.04$$

$$\text{and } n_{final} = \frac{n}{1 + \frac{n}{N}}$$

$$= \frac{96.04}{1 + \frac{96.04}{130}}$$

$$= 55.23$$

Note. [a]The formula used assumes simple random sampling, but we actually used stratified random sampling. The proper formulas are more complicated, but as a practical matter, in this instance the answers will not be much different; whatever small differences there are will likely be in our favor.

Part II. Calculations for Confidence and Precision for Pilot Sample Size[a]

These calculations were used for estimating confidence and precision, based on the actual pilot sample size; we wanted to determine how the small sample would alter our confidence and/or precision, given

$$nfinal = \frac{N\,t^2\,p\,q}{(N-1)\,d^2 + t^2\,p\,q}$$

where N = 130 (sampling frame size)

n = 30 (actual sample size)

p = .5 ("worst case" proportion to have ≥ 1 major methodological error))

q = 1 - p = .5 (for calculating proportion variance)

precision = d or confidence (from t)

precision = d	or	confidence (from t)
$d2 = \frac{(N-n^{final})\,t^2\,p\,q}{(N-1)\,n^{final}}$		$t^2 = \frac{d^2\,(N-1)\,n^{final}}{(N-n^{final})\,p\,q}$
where t =1.96		where d = .1
then d^2 = .0258		then t^2 = 1.548
d = .1575		t = 1.244
		confidence = 78.7% [Remington & Schork (1985); Table A-4, Column C, p. 36.

Note. [a]The formula used assumes simple random sampling, but we actually used stratified random sampling. The proper formulas are more complicated, but as a practical matter, in this instance the answers will not be much different; whatever small differences there are will likely be in our favor.

Appendix E

Institutional Review Board Letter

THE UNIVERSITY OF NORTH CAROLINA
AT
CHAPEL HILL

Department of Health Policy and Administration
School of Public Health

The University of North Carolina at Chapel Hill
Kron Building 514A
Chapel Hill, N.C. 27514-6201

May 4, 1987

Dr. Maija L. Selby
Department of Public Health Nursing
School of Public Health 201 H
The University of North Carolina
Chapel Hill, NC 27514

Dear Dr. Selby:

Your proposal entitled "Evaluation of Methodology, Statistical Analysis, and Reporting in Published Nursing Research" is exempt from IRB review. Exemption is claimed on number 3 of the criteria for exemption outlined in Federal Regulations 45 CFR Part 46: Federal Register 46(16), dated 1/26/81.

This exemption should be noted on the HHS-596 Form, if applicable, and must include the statement "Exemption is claimed on number __." The number claimed is designated above.

Sincerely,

Arnold D. Kaluzny, Ph.D.
Chairman
Institutional Review Board
on Research Involving
Human Subjects

ADK:dfc

Appendix F

Calculations for a 95% Confidence Interval for the Proportion of Articles With a Major Error in Sampling Methodology

Appendix F

Calculations for a 95% Confidence Interval for the Proportion of Articles With a Major Error in Sampling Methodology

These calculations were used for estimating the confidence interval, based on the the actual pilot sample size, given

$$\hat{p} \pm \sqrt{\frac{t2\ \hat{p}\hat{q}}{n}\frac{(N-n)}{(N-1)}}$$

where N = 130 (sampling frame size)

n = 30 (final sample size)

$\hat{p}$ = .967 (estimated from sample data)

$\hat{q}$ = .033 (estimated from sample data)

t = 1.96 (for 95% confidence)

$.967 \pm \sqrt{.0031677}$

$.967 \pm .056$ (.911, 1.023)

Therefore, the upper and lower limits of the 95% confidence interval were .91 (91%) and 1.0 (100%).